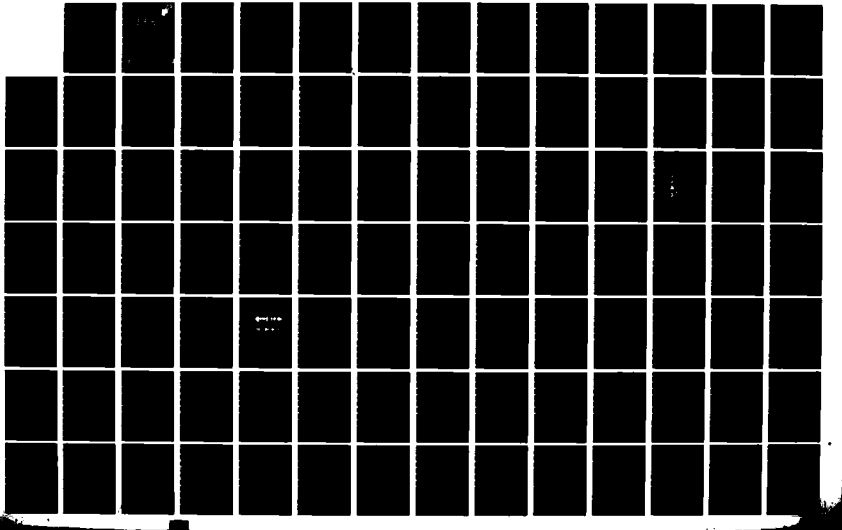
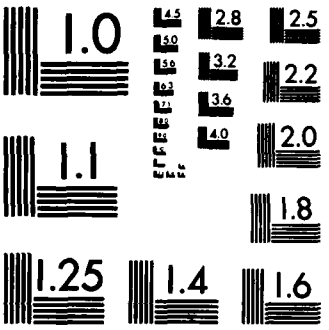
AD-A135 702 PROCESSING TECHNIQUES FOR INTELLIGIBILITY IMPROVEMENT 1/2
TO SPEECH WITH CO-C..(U) SIGNAL TECHNOLOGY INC GOLETA
CA B A HANSON ET AL. SEP 83 RADC-TR-83-225
UNCLASSIFIED F30602-81-C-0226 F/G 17/2 NL





RADC-TR-83-225
Final Technical Report
September 1983

PROCESSING TECHNIQUES FOR INTELLIGIBILITY IMPROVEMENT TO SPEECH WITH CO-CHANNEL INTERFERENCE

Signal Technology, Inc.

Brian A. Hanson and David Y. Wong

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

ROME AIR DEVELOPMENT CENTER
Air Force Systems Command
Griffiss Air Force Base, NY 13441

This report has been reviewed by the RADC Public Affairs Office (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

RADC-TR-83-225 has been reviewed and is approved for publication.

APPROVED: Melvin G Manor Jr

MELVIN G. MANOR, JR.
Project Engineer

APPROVED: Thadeus J Domurat

THADEUS J. DOMURAT
Acting Technical Director
Intelligence & Reconnaissance Division

FOR THE COMMANDER: Donald A Brantingham

DONALD A. BRANTINGHAM
Plans Office

If your address has changed or if you wish to be removed from the RADC mailing list, or if the addressee is no longer employed by your organization, please notify RADC (IRAA) Griffiss AFB NY 13441. This will assist us in maintaining a current mailing list.

Do not return copies of this report unless contractual obligations or notices on a specific document requires that it be returned.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER RADC-TR-83-225	2. GOVT ACCESSION NO. A135702 / 3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) PROCESSING TECHNIQUES FOR INTELLIGIBILITY IMPROVEMENT TO SPEECH WITH CO-CHANNEL INTERFERENCE	5. TYPE OF REPORT & PERIOD COVERED Final Technical Report June 81 - August 83 6. PERFORMING ORG. REPORT NUMBER N/A
7. AUTHOR(s) Brian A. Hanson David Y. Wong	8. CONTRACT OR GRANT NUMBER(s) F30602-81-C-0226
9. PERFORMING ORGANIZATION NAME AND ADDRESS Signal Technology, Inc. 5951 Encina Road Goleta CA 93117 *	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 35885G 7055A742
11. CONTROLLING OFFICE NAME AND ADDRESS Rome Air Development Center (IRAA) Griffiss AFB NY 13441	12. REPORT DATE September 1983 13. NUMBER OF PAGES 138
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) Same	15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE N/A

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

Same

18. SUPPLEMENTARY NOTES

RADC Project Engineer: Melvin G. Manor (IRAA)

* Subcontractor: Speech Technology Lab
3888 State Street
Santa Barbara CA 93105

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Speech Enhancement	Co-Channel Interference	Linear Prediction (LPC)
Speech Intelligibility	Voice Interference	Harmonic Synthesis
Spectral Subtraction	Speech Processing	Spectral Sampling
Co-Channel Separation	Post-Processing	Spectral Distortion Measure

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Enhancing speech that has been corrupted during transmission or recording is a widely researched problem. A special case is when the interference is a second talker's voice. This report summarizes research on processing techniques that improve the intelligibility of the desired speech signal in the presence of such voice interference.

Formal subjective intelligibility test procedures for evaluating enhancement algorithms for the voice interference problem are first developed.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Computational objective measures based on spectral criteria are also developed to provide testing during intermediate stages of algorithm development.

The first algorithm presented estimates spectral parameters of the desired speech by harmonic sampling and regenerates this speech by a novel synthesis technique. Although limited-scale testing on this "harmonic extraction" algorithm does not show any intelligibility improvement, analysis of the results leads to several new research directions. The most important conclusions are that negative decibel signal-to-noise ratio (SNR) voice-interfered speech has the greatest potential for intelligibility improvement, and that interference suppression is the logical choice over signal extraction for usch SNR conditions.

Suppression of the voiced segments of the interfering speech is considered next. There are two parts to this approach, interference estimation and removal. Based on analytical and experimental findings, spectral magnitude subtraction is selected as the interference removal technique. Three interference estimation methods are developed and compared using this spectral subtraction technique. The best of the three, harmonic magnitude suppression (HMS), is evaluated with formal intelligibility testing.

The formal intelligibility testing on the HMS algorithm shows intelligibility improvement at a 98% confidence level for speech with voice interference at -12 dB SNR. No previous work on this voice-interference problem has reported such measured intelligibility gains. Closer analysis of the test data also shows that the intelligibility improvement tends to increase with decreasing SNR. These results indicate that HMS processing representa a major step towards realization of a complete intelligibility enhancement system.

TABLE OF CONTENTS

1.0 INTRODUCTION

1.1 Problem Definition

A recurrent problem in the transmission and recording of speech signals is the crosstalk between communication channels. For example, much effort has gone into analyzing and avoiding such interference in parallel telephone circuits. Where feasible, the preventive approach is the best for solving the crosstalk problem. However, this is not always possible due to different operational situations. There is thus a strong interest in signal processing techniques for separating two voices which exist in a single channel. This will be referred to in this report as the "co-channel separation" problem.

The purpose of this research is to develop post-processing techniques for co-channel separation. In speech enhancement research, the goal varies from improving signal-to-noise ratio (SNR) to enhancing the quality or listenability, to improving intelligibility. While a number of claims have been made on quality or SNR improvements, no research to date has been able to demonstrate any measurable improvement in the intelligibility of the speech after co-channel separation processing. Enhancement of the intelligibility of the desired voice signal (which has been interfered by a second voice) is the ultimate concern in this

study. Even though other attributes are important, the transmission of information from the speaker to the listener through the communication system is the primary goal; thus intelligibility of the received speech is the most significant measure of system performance. In fact, the secondary goals of reducing fatigue and improving "listenability" [Berouti et al. 1979] often follow as a natural consequence of intelligibility improvement.

The basic problem definition of this study is summarized in Fig. 1-1. The received signal is the sum of two speech signals produced by two talkers. Although there are also multiple speaker situations of interest, only two talkers are considered in this study, both for simplicity and because this is the most commonly encountered situation. One of the two voices (s_1) will be denoted the "desired signal" or speech, and the other (s_2) is the "interfering noise". The input of the system developed in this study is s_1+s_2, and the output is an enhanced version (or estimate) of the desired talker's speech, $\hat{s}_1$.

In this study no other information is assumed available to the co-channel separation algorithms besides the summed speech signal. This assumption considerably constrains the approaches that can be taken. For example, if large amounts of a priori data are available from either the desired or interfering speakers alone, then certain speaker charac-

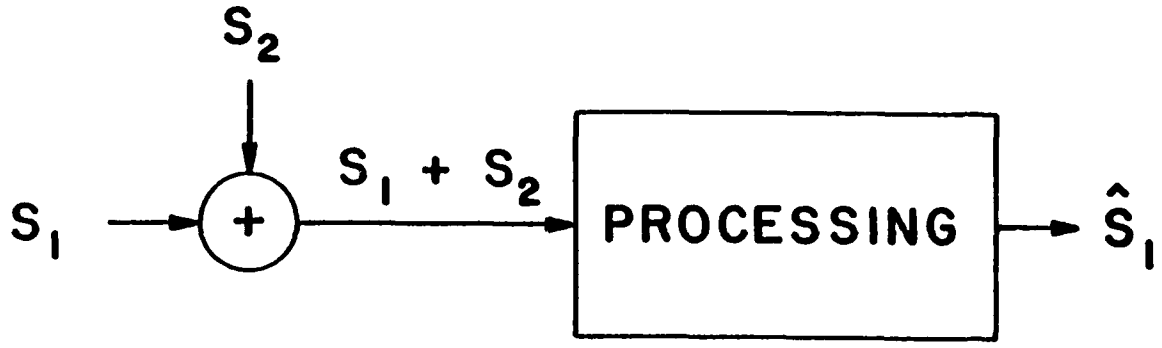


Fig. 1-1: Co-Channel Separation Problem

teristics can be identified and used in the separation process. Or, if supplementary data were available simultaneously with the co-channel speech, such as reference signals which are correlated with either the signal or noise, then adaptive noise cancellation techniques could be applied [Strube 1981]. Also, because the co-channel speech, s_1+s_2, is monophonic (i.e. single-channel), binaural listening techniques [see e.g. Berlin and McNeil 1976] are of little use.

The problem definition as described above is listed in Table 1-1. It should be noted that this problem definition is representative of many practical situations.

- Input signal is monophonic, with
- one desired voice and
- one additive interfering voice.
- No a priori individual speaker information, training data sets, or signal or noise references available.
- The goal is to develop post-processing techniques to
- enhance the intelligibility of the desired voice.

Table 1-1 Problem Definition

1.2 Review of Previous Research

Although a considerable amount of research has been done on enhancement of speech in the presence of various types of noise and distortion (see e.g. [Lim 1983]), only a limited number of these studies have been concerned with the co-channel separation problem. This section briefly summarizes the previous studies on this subject.

A technique for co-channel separation that attempt to filter out all spectral components of the co-channel except those around the pitch harmonic frequencies of the desired speaker was suggested by Shields [1970]. This "comb-filtering" technique was implemented in the time domain and made adaptive to changes in pitch frequency by Frazier [1975]. Comprehensive testing of Frazier's technique was conducted by Perlmutter et al. [1977] for different lengths of the comb filter. Some of Perlmutter's better results are shown in Fig. 1-2. The intelligibility of the desired speech after processing was found to be always less than in the original unprocessed co-channel signal; also as the length of the comb filter increased, the intelligibility usually decreased even further. Two different methods of handling the unvoiced (i.e. non-periodic) segments were also evaluated. In the attenuation technique, the unvoiced segments are simply reduced by a constant amount and passed directly to the output. For the inertial

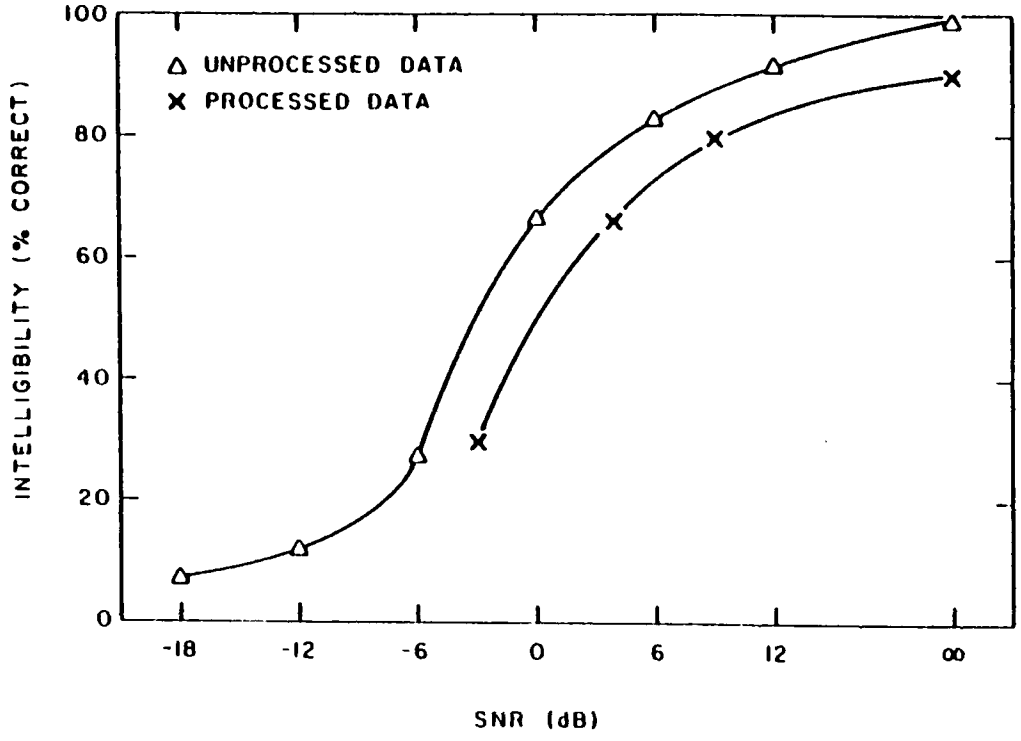


Fig. 1-2: Intelligibility vs. SNR (sketched from [Perlmutter et al. 1977])

method, the comb filtering is continued into the unvoiced desired speech segments using the last pitch value calculated for the preceding voiced speech. While both methods failed to yield improved intelligibility over the unprocessed data, it is interesting to note that the attenuation method generally provided better results than inertial unvoiced processing.

In Perlmutter's experiments the pitch contour used by the separation algorithm is extracted from the individual speech data before the speech is combined to form the co-channel signal. Although this procedure is obviously not applicable for actual operation, where only the co-channel signal is available, this experimental methodology allows one to divide the co-channel separation problem into two subproblems: i) pitch detection on co-channel speech and ii) desired speaker enhancement processing. This division allows the enhancement processing to be considered alone; once this problem is adequately solved, the co-channel pitch issue can be tackled. The same methodology is adopted in this study.

Other pitch-based separation approaches have been reported by Dick [1980], Everton [1975], Parsons and Weiss [1975], and Parsons [1975, 1976, 1978, 1979]. These can be divided into time domain techniques (e.g. Frazier's comb filtering described earlier) or frequency domain methods.

The research reported by Parsons is typical of the frequency domain methods, so his work will be discussed here. The basic procedure, as presented in [Parsons and Weiss 1975, Parsons 1975,1976], is a frequency domain technique which combines pitch detection and desired speaker enhancement into a single algorithm.

Parsons' algorithm starts with estimation of the frequency, amplitude, and phase for each peak in a short term spectrum. This peak information is used to estimate the pitch of the desired and interfering speech, which in turn allows each peak to be assigned to one of the speakers (after all overlapping peaks have been resolved with additional processing). Once the peak assignment is completed, Parsons' procedure selects resynthesis of either the desired or the interfering speaker spectra. When the interference is synthesized, Parsons subtracts it from the original co-channel signal to obtain the desired speech. He reports that the subtraction results were not satisfactory and concentrates subsequent efforts on direct synthesis of the desired speech. Although the synthesis approach is reported to provide "fair to excellent" speech intelligibility, no formal intelligibility testing has been reported.

An interesting departure from the pitch-based approach is the work of Young and Goodman [1977]. They suggest that peak clipping of the pre-whitened co-channel speech may

improve the intelligibility of the desired speaker. This is based on the well known fact that clipping does not seriously affect single-speaker intelligibility (see e.g. [Martin 1950]). The assumption is that in cases where the desired speech is weaker than the interference, clipping will equalize the energies of the desired and interfering speech, thereby improving intelligibility. Young and Goodman ran tests on this concept using co-channel data with five simultaneously interfering speakers. However the test results indicate that the intelligibility of the desired speaker is severely reduced by the prewhitening/clipping processing.

All past studies have failed to demonstrate measurable intelligibility gains. At the onset of this study, it is clear that there is serious doubt that any signal processing technique can improve the intelligibility of co-channel interfered speech.

1.3 Outline of Report

One of the key steps in developing a co-channel separation system is evaluating the results. The formulation of a well-defined method for formal subjective intelligibility evaluation is developed in this study. While the subjective measure is the preferred criterion, the test procedure is extremely time consuming. Therefore, computational objec-

tive performance measures are developed for preliminary screening and evaluation. The details of the measures, subjective and objective evaluation methods, are discussed in chapter two.

Several different approaches to co-channel separation are investigated. The first approach is to estimate and extract the desired signal, based on a harmonic synthesis technique. Details of this signal extraction approach are discussed in chapter three. Preliminary testing performed on this extraction system is also reported. Although the tests on this extraction system indicate no intelligibility gains, the results provide new insights into the problem which lead to the second approach.

The second approach to co-channel separation is to estimate and then remove or suppress the interference signal. The development starts with the selection of an appropriate spectral subtraction algorithm. To apply spectral subtraction to the co-channel problem, the interference spectrum must be estimated. Hence an estimation approach is developed. Details of these studies are discussed in chapter four.

Subjective tests on the spectral suppression technique are performed. The test results demonstrate that for low SNR co-channel speech, a statistically significant intelligibility gain is realized with the proposed post-processing

technique. Details of the test are presented in chapter five.

Conclusions of this research and recommendations for future research into implementing a total co-channel separation system are presented in the last chapter of this report, chapter six.

2.0 ALGORITHM PERFORMANCE MEASURES

Before the development of a co-channel separation algorithm, it is important to first define how the processing algorithms can be evaluated. This chapter discusses two different approaches to the performance evaluation problem. The first is subjective listening tests. A formal procedure for this is discussed in section 2.1. The second technique, discussed in section 2.2, is calculation of numerical measures that approximate the behavior of human auditory preprocessing, which is correlated to intelligibility.

2.1 Formal Intelligibility Testing

This section covers the procedures used in the intelligibility tests. Deviations from these general procedures, and the particular parameters used in each test (i.e. number of subjects, SNR's, etc .), are discussed in subsequent chapters.

Test Objectives

A number of formal subjective testing procedures have been developed for both speech quality and intelligibility evaluation [IEEE 1969, Hawley 1977]. These procedures were first developed for testing speech therapy subjects, they were later developed for evaluating communications systems, and more recently they are even used for testing electronic

voice synthesizers. The goals of these test procedures are to reliably and meaningfully quantify the quality or intelligibility of speech. For intelligibility testing, the best known procedures are the modified rhyme test [House et al. 1965] and the diagnostic rhyme test [Voiers 1977]. While these procedures are well designed and quite widely adopted by speech therapists and engineers alike, they are not appropriate for this research because the test material consists of isolated rhyme words. In order to properly simulate a realistic co-channel interference situation, continuous speech data is necessary, requiring new and different test procedures.

There has been only one other published report of intelligibility testing for the co-channel separation problem with a single interfering speaker [Perlmutter et al. 1977]. Some of the procedures developed in the present study are derived from this earlier work. However, due to differences in the research application and emphasis, important departures are necessary. The intelligibility test procedures developed in this study are discussed below.

Test Material

The first step in the intelligibility testing procedure is the collection and preparation of a data base which is representative of the data encountered by the co-channel separation system. Earlier testing in this area [Perlmutter

et al. 1977] used "syntactically normal nonsense sentences" for the desired speaker. These consisted of a fixed pattern of verb, adjective, and nouns (e.g. "The round work came the well"). The interference signals were sentences from the "1965 Revised List of Phonetically Balanced Sentences" [Appendix C of IEEE 1969]. The use of nonsense sentences for the desired (target) signal is to eliminate variabilities due to linguistic cues above the syntactical level. The use of PB sentences as interference eased the problem of "target-jammer alignment" (i.e. this avoided identical speech-pause patterns in the target and jammer). We feel that the artificial nonsense sentences are unnecessary, and in fact unrealistic, so in the testing procedure used in this study phonetically balanced sentences are used for both the desired and interference speech. This use of meaningful sentences allows the listeners to make full use of all levels of linguistic cues for both the signal and interference, providing a more realistic test for the system.

Test Data

The test material (PB sentences) was read by a panel of (two or more) speakers. These readings were recorded on audiotape and then digitized at 10 kHz, 16 bits/sample. The input test data was generated by summing the speech from two of the speakers at the specified signal-to-noise ratio (SNR). This SNR is defined as the ratio of the average

energy in only the speech portions of the desired and interfering signals; pause segments are not included in the averages. The "pause or speech" decision is made by measuring the background noise level just before the start of the utterance, and then using this energy value as a threshold to detect pause segments. Thus the SNR can be written as the ratio of the sums of the energies from the thresholded signal and noise speech frames:

$$SNR = \frac{\frac{1}{N_s} \sum_i g_T\,[\text{signal energy}(i)]}{\frac{1}{N_n} \sum_i g_T\,[\text{noise energy}(i)]} \qquad (2\text{-}1)$$

where

energy(i) = energy evaluated for i-th (20 msec) frame

$$g_T[x] = \begin{cases} 0 & \text{for } x<T \\ x & \text{for } x \geq T \end{cases}$$

T = pause energy threshold
N_s = number of signal frames above threshold
N_n = number of noise frames above threshold

Pause removal before SNR computation is also adopted in speech coding research to generate "segmental SNR" [Jayant and Noll 1982, Noll 1974].

Another important consideration in test material preparation is the alignment of the desired and interfering speech signals. The sentences used for the desired speech and the interference are first sorted according to duration. The longest interfering speech segment is mixed with the longest desired speech data and so on. The interference

signal is generally centered with respect to the desired speech signal, leading to maximum coverage of the desired speech by the interference. Perfect synchronization of signal and interference is neither practical nor desirable as the speech-pause pattern of two voices on different channels will not likely be synchronized. While the overlap is somewhat maximized by proper alignment, the exact pattern of overlap is left to chance to approximate a realistic co-channel situation. Variability is reduced by including a large enough set ($\geq$ 10) of PB sentences. The data described forms the input or "unprocessed" data. After passing this data through the speech enhancement algorithm under consideration, the output forms a second set of data, the "processed" data.

Listening Panel

A panel of subjects is recruited to compare the intelligibility of the processed versus the unprocessed data. In order to avoid possible retention effects from previously heard speech, subjects chosen for the listening panel are completely unfamiliar with the text of the speech data used in the intelligibility tests. Most of the listeners are professionals or graduate students in the speech and hearing (or linguistic) field. Such "experienced listeners" are selected because it is thought that they will be well-motivated and hence more consistent in performance. This

expectation is generally verified in comparing their results to those of the less experienced listeners. Several "less experienced" listeners were included in the panel to provide enough data to get statistically significant results.

Test Session

Listening to processed and unprocessed co-channel data is conducted in individual sessions for each listener. Two listening procedures are used. In the first procedure, the "comparison" test, half of the data presented to the listening subject in a session is unprocessed and the other half is processed. The processed and unprocessed data are different sentences spoken by the same speakers. The speech data presented to the listeners are arranged so that half of the subjects hear a particular sentence in its unprocessed form and the other half of the subjects hear it as processed. A simple case for this type of test with just two sentences and two subjects is illustrated in Table 2-1(a). With a sufficiently large panel of subjects and sentences, the variability due to subjects and test material is averaged out.

SUBJECT A hears:	U1 and P2
SUBJECT B hears:	P1 and U2

(a) Intelligibility comparison test presentation

SUBJECT A hears:	U1 then U1 and U2 then P2
SUBJECT B hears:	U1 then P1 and U2 then U2

(b) Intelligibility improvement test presentation

Table 2-1: Intelligibility Testing Techniques
(U1=unprocessed sentence 1 and P1=processed sentence 1)

The second test procedure evaluates the degree to which the processed data adds to (or improves) the intelligibility of the unprocessed data. The procedure is the same as that above except that both the processed and the unprocessed data for half of the sentences are presented to the listeners. The other half of the test material is presented as unprocessed only (to give an equal number of repetitions of the data, the unprocessed-only data is repeated twice). A simple case of such an "intelligibility improvement" test is indicated in Table 2-1(b).

The comparison testing technique compares the intelligibility of the processed versus the unprocessed data, while improvement testing determines whether the processing improves the intelligibility of the input co-channel data. The choice of intelligibility testing method to be used is determined by how the enhancement algorithm is used. When

the algorithm in chapter three was developed, it was thought that unprocessed speech would be completely replaced by processed speech, so the comparison test procedure was used. The results of the test, however, showed that unprocessed speech quite often is very intelligible, hence it is desirable to keep the unprocessed data where possible. The improvement testing procedure is the preferred method in such cases where the original unprocessed co-channel signal is assumed to be also available.

At the start of a test session, the subject receives written instructions for the test. A copy of these instructions is included in appendix A. The subject's task is to orthographically transcribe as many of the intelligible words as possible (including guesses) from all of the presented data. To avoid biasing the subjects, the nature of the research project is not discussed until after the session is completed. This provides a uniform understanding of the test for each subject.

The listener is then seated in a sound booth to avoid possible outside noise interference or interruptions. The booth is equipped with a D/A port, headphone amplifier, and computer terminal. A short demonstration of the interactive listening program (used by the listener to control the playback of speech samples in the test) is run to familiarize the subject with its operation. The subject is then left to

proceed at his own pace through the test material with the interactive procedure.

The subjects are allowed as many repeats of the material as needed to complete the transcription (multiple repeats are used to determine the maximum amount of intelligible information in the unprocessed and processed speech).

Scoring

The rules used for scoring the subjects' transcriptions are listed in Table 2-2. The primary goal of evaluating intelligibility improvement implies that the semantic information (i.e. meaning) of each utterance is most important, and the scoring rules are based on this assumption. The only exception is that homonyms are accepted as correct because, for the low intelligibility cases dealt with in this study, the contextual and grammatical clues are not always present to select the right homonym. For example, if the only intelligible word in a phrase is "to", the responses "too" or "two" are scored as correct.

In the testing procedure used by Perlmutter et al. [1977], "perfect" transcription of each word was required. In the present study, partial score rules are set up for transcribed words that are very close to the correct text, as shown in rule two. The rule allows for the insertion, deletion, or substitution of one prefix or suffix mor-

pheme. An example is allocation of one-half point for transcribing "burn" or "burns" when the spoken word is "burned." Such morphemic errors are allowed because the semantic information is generally preserved.

Multiple guesses are also allowed as described by rule three. For example, if two responses ("fired" or "tired") are transcribed when the correct word is "tired", one-half point is given. Finally, the score multiplication of rule 4 handles cases that involve both scoring rules 2 and 3 (i.e. multiple guesses where one of the responses is very close, as defined by rule 2).

1) One point for perfect word (or homonym).

2) One-half point for word with correct root morpheme (or homonym) with incorrect prefix or suffix morpheme which is only a single phoneme in duration. For example, adding an "s" for a plural or making a tense change with an added "ed".

3) 1/N point for one of N responses correct.

4) Rules are multiplicative (e.g. if one of two choices satisfies rule #2 above, then score is 1/4).

Table 2-2 Scoring Rules

2.2 Computational Objective Performance Measures

Formal subjective intelligibility testing as described in section 2.1 is time consuming because many subjects and

test samples are required to obtain statistically significant results. Testing at all stages of algorithm development is thus not practical. Therefore a computational objective measure that is correlated with intelligibility is needed for testing intermediate co-channel separation algorithmic choices.

Signal-to-noise ratio has been shown to be correlated with intelligibility for laboratory generated unprocessed co-channel data [Miller 1947, Perlmutter et al. 1977]. One disadvantage of using SNR for evaluating the intelligibility of processed co-channel speech is the equal weighting given to all frequencies in calculating SNR. The co-channel separation processing may eliminate the interference only in part of the frequency spectrum, and the effects of the remaining interference are highly frequency dependent (i.e. the interference in one part of the frequency spectrum may contribute to the loss in intelligibility much more than the interference in another part of the spectrum). Evaluation of these frequency-dependent effects requires consideration of several aspects of human auditory pre-processing.

Numerous psychoacoustic experiments have been conducted to study the effects of interference on human auditory perception (see e.g. [Small 1973, Harris 1974, Gelfand 1981]). An important conclusion of these studies is that the initial stage of auditory processing has characteristics similar to

a bank of bandpass filters. These bandpass characteristics define the manner and frequency ranges (known as critical bands) over which auditory stimuli interact. Scharf [1970] summarizes much of the work in this field, and his graph of critical bandwidths versus frequency is shown by the solid curve in Fig. 2-1. The so-called "Bark" scale [Zwicker 1961] approximates this curve by modifying the frequency axis so that the critical bandwidth is constant (i.e. one Bark) everywhere on the scale. An approximate expression given by Fourcin et al. [1977] relating frequency (in Hz) to Barks (z) is:

$$f = 600 \sinh(z/6) \tag{2-2}$$

Comparison of the Bark scale to the well known mel scale shows that these two scales are quite similar.

Filtering functions (i.e. magnitude responses) which model the observed psychoacoustical bandpass characteristics are given by Schroeder in [Fourcin et al. 1977]. An improved version of Schroeder's function, proposed by Sekey and Hanson [1983], is used in the present work. Expressed in Barks this function is:

$$10\text{Log}F(z) = 7.0 - 7.5(z-0.215) - 17.5[0.196+(z-0.215)^2]^{1/2} \tag{2-3}$$

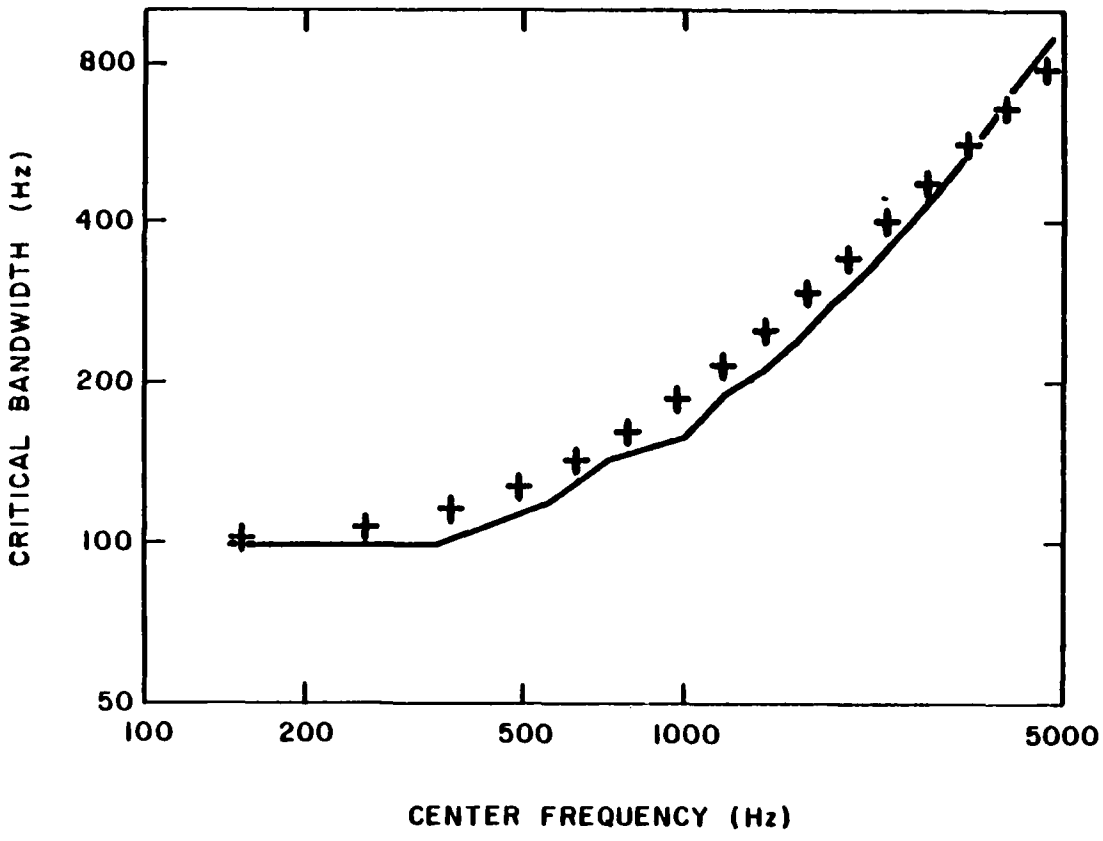


Fig. 2-1: Critical Bandwidth vs. Center Frequency of Critical Band (curve from Scharf's [1970] tabulated values and crosses from [Sekey and Hanson 1983])

Using equations (2-2) and (2-3) a set of sixteen filter functions can be derived which cover the frequency range of interest in this study (100 to 5000 Hz), with adjacent functions crossing approximately at their 3 dB points. These sixteen filter functions are plotted in Fig. 2-2. The bandwidths of these filter functions, indicated by the crosses on Fig. 2-1, generally agree with the bandwidths given by Scharf.

A SNR-type measure which uses critical band filters similar to the above is the well-known articulation index (AI). The AI, as defined in [Kryter 1962a, ANSI 1969], is basically an average of the SNR's from each critical band. An important step in AI calculation is to assure that the SNR from each frequency band does not exceed a certain maximum (or minimum) value. This SNR limiting implies that increases in a critical band's SNR do not increase the AI (and by implication the intelligibility) once the SNR exceeds a maximum value; similarly, a critical band's contribution to the AI (and intelligibility) does not decrease further as the SNR drops below a minimum. The validity of this procedure is supported by experimental intelligibility data (e.g. Fig. 1-2). Kryter [1962a] uses limits of 30 and 0 dB in his formulation of the AI. However, in co-channel speech different limits are recommended. Perlmutter et al. [1977] demonstrated that the intelligibility of co-channel

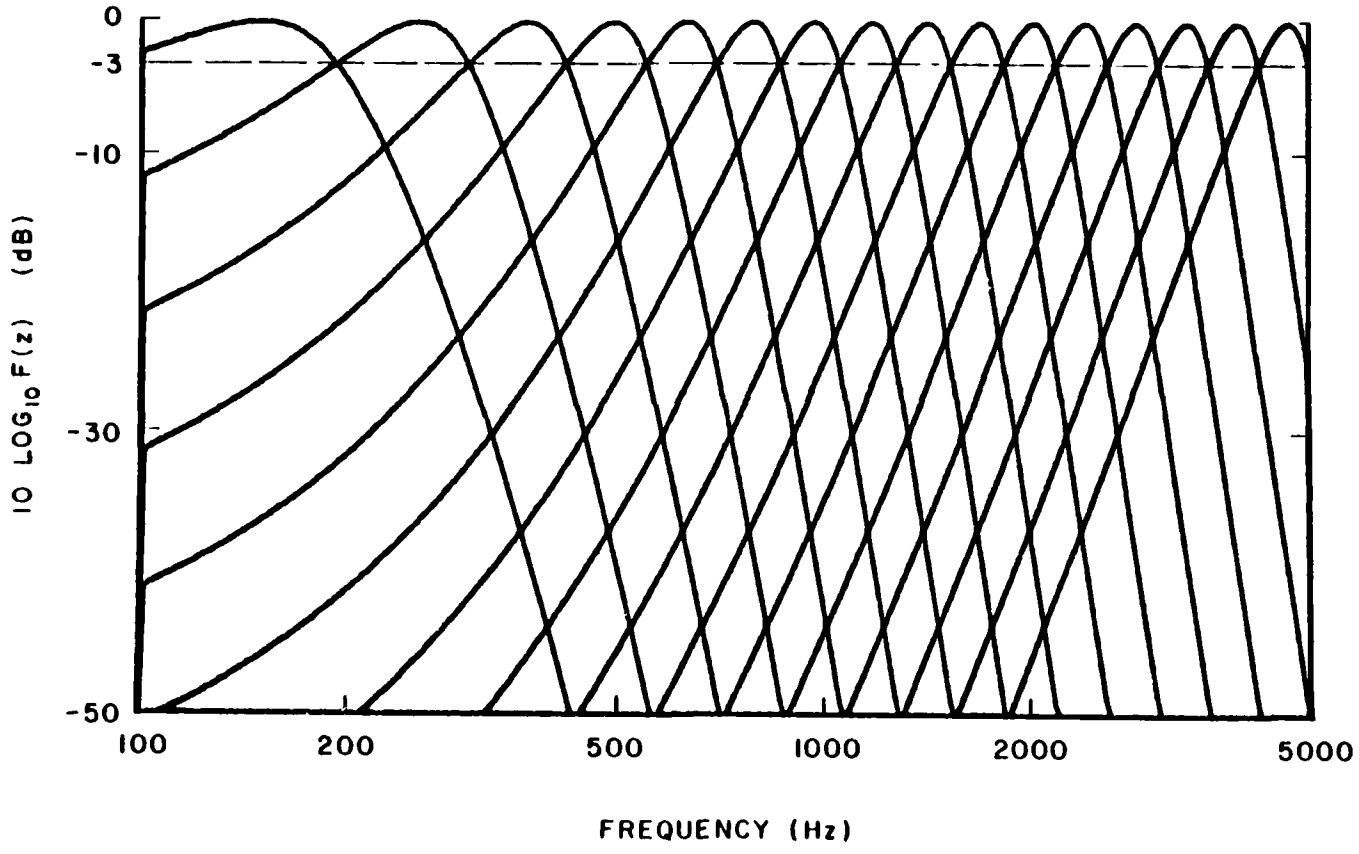


Fig. 2-2: Critical Band Filtering Functions vs. Frequency

speech varies between about 10% and 90% over a range of SNR's from -18 dB to +12 dB, with a monotonic increase in intelligibility with SNR between these extremes (see Fig. 1-2). Thus, for co-channel speech +12 and -18 dB are more appropriate SNR limits.

The articulation index has been shown to be correlated to intelligibility of noisy speech in numerous situations (see e.g. [Kryter 1962b]). Unfortunately, when AI (or any SNR-based measure) is used to evaluate processed co-channel speech, it is not always correlated with intelligibility. This problem arises because calculation of the SNR values used in the AI requires an estimate of the noise remaining after separation processing. This noise estimate, and the resulting AI or SNR, can be seriously affected by separation-processing-induced distortions (e.g. phase delays) which have little effect on intelligibility. Thus, it is necessary to develop a computational measure which incorporates the psychoacoustical aspects of the AI discussed above, but does not require an estimate of the noise remaining after co-channel separation processing.

A measure for evaluating intelligibility that does not require noise estimates is the spectral distortion measure (SDM). A number of these measures have been developed for speech coding research [Gray et al. 1980, Gray and Markel 1976]. Recently Boll and Wahlford [1983] also applied SDM's

to wideband noise reduction research. In the rest of this section the mathematical definition and properties of one class of SDM are reviewed, and several concepts from the AI are used to develop a modified SDM for co-channel algorithm evaluation. Examples of the calculation and application of this SDM will also be presented.

Spectral distortion measures are used to evaluate co-channel separation algorithms in development work by comparing SDM's between the clean desired speech and the co-channel speech before and after processing. The class of SDM's considered in this work measures the degree to which the co-channel speech log spectrum matches the log spectrum of the desired speech. The perceptual basis behind such measures is that the closeness of spectral matching expressed by the SDM correlates with intelligibility.

The SDM of interest is the mean absolute log SDM. It is a typical SDM technique calculated by taking log spectral differences at each frequency and integrating these over the whole frequency band. Taking the p-th power of the difference and using discrete spectra, the general log difference SDM is defined by:

$$SDM_p = \left\{ \sum_{k=1}^{K} \left| LOG|S(k)| - LOG|\hat{S}(k)| \right|^p \right\}^{\frac{1}{p}} \qquad (2\text{-}4)$$

where

$S(k)$, $\hat{S}(k)$ = K point DFT's of desired and co-channel speech, respectively

The value of p in equation (2-4) controls the relative weighting of large and small spectral differences between the desired and co-channel speech. For example, as p approaches infinity the value of SDM_p becomes dependent only on the peak spectral difference. The mean absolute log case (p=1) calculates the area between the two log spectra, with all spectral differences weighted equally. This SDM with p=1 is an interesting case since, as the noise becomes considerably larger than the signal in energy (i.e. SNR << 0 dB), the SDM approaches the negative of the logarithmic average SNR. This is shown in the following, where $S(k)$, $\hat{S}(k)$, and $\hat{N}(k)$ represent the discrete spectra of the desired speech, co-channel signal, and co-channel interference, respectively:

$$SDM_{p=1} = \sum_{k=1}^{K} \left| LOG \frac{|S(k)|}{|\hat{S}(k)|} \right| = \sum_{k=1}^{K} \left| LOG \frac{|S(k)|}{|S(k)+\hat{N}(k)|} \right| \qquad (2\text{-}5)$$

If SNR << 0 dB, then $|S(k)| << |\hat{N}(k)|$ and:

$$SDM_{p=1} \approx \sum_{k=1}^{K} \left| LOG \frac{|S(k)|}{|\hat{N}(k)|} \right| = -\text{average SNR} \qquad (2\text{-}6)$$

Rather than directly summing the spectral differences over all frequencies as in equation (2-4), critical band weighting can be incorporated into the SDM, as in the AI calculation. This is achieved by calculating critical band power outputs for the desired and co-channel speech, and then taking log differences. These operations are indicated below for the p=1 case:

$$SDM_{cb} = \sum_{i=1}^{16} \left| 10LOG_{10} \frac{pwr_i(s)}{pwr_i(\hat{s})} \right| \qquad (2\text{-}7)$$

where

$pwr_i(\ )$ = power calculated in i-th critical band (for desired or co-channel speech signals)

Use of critical band filtering outputs in a SDM, as in equation (2-7), has been considered before in other areas of speech research, such as [Davis and Mermelstein 1980]. As in the derivation of equation (2-6), it can be shown that as the SNR decreases, SDM_{cb} becomes roughly proportional to the negative AI. Thus the SDM_{cb} incorporates some properties of the AI without having the computational difficulties of the AI for processed data (i.e. estimation of the noise).

Another feature of AI calculation incorporated in SDM_{cb} is the SNR limiting imposed within each individual frequency band. In the AI calculation, the SNR for each band is limited to a certain maximum value because it is assumed that

when the maximum SNR is reached, increasing the SNR further does little to increase intelligibility. This peak SNR clipping property is approximated by the log differences in the SDM, which contribute little to the total SDM whenever the powers of $\hat{s}$ and s in a critical band are close. For a lower SNR limit, the value of -18 dB was suggested earlier for use in the AI; since log power differences of s and $\hat{s}$ approach the negative SNR for low SNR values, this -18 dB lower limit on SNR can be approximated by limiting the log spectral differences in equation (2-7) at a maximum of +18 dB. Because this limit tends to emphasize the less distorted parts of the processed speech, both SDM's with and without the +18 dB limit will be calculated for comparison in most cases (the limited SDM's values will be labeled as "18 dB limited").

Speech spectra are relatively invariant only over short time intervals (typically less than 40 msec), so the SDM of equation (2-7) is evaluated for short time segments of the co-channel data and original clean desired speech. A typical short-term SDM contour is shown in Fig. 2-3. The SDM's for a co-channel signal before and after processing with a separation algorithm are calculated every 20 msec and plotted versus time below the signal (i.e. desired speech) and noise waveforms. The SDM contour shows where the separation algorithm improves the spectral match with the clean desired

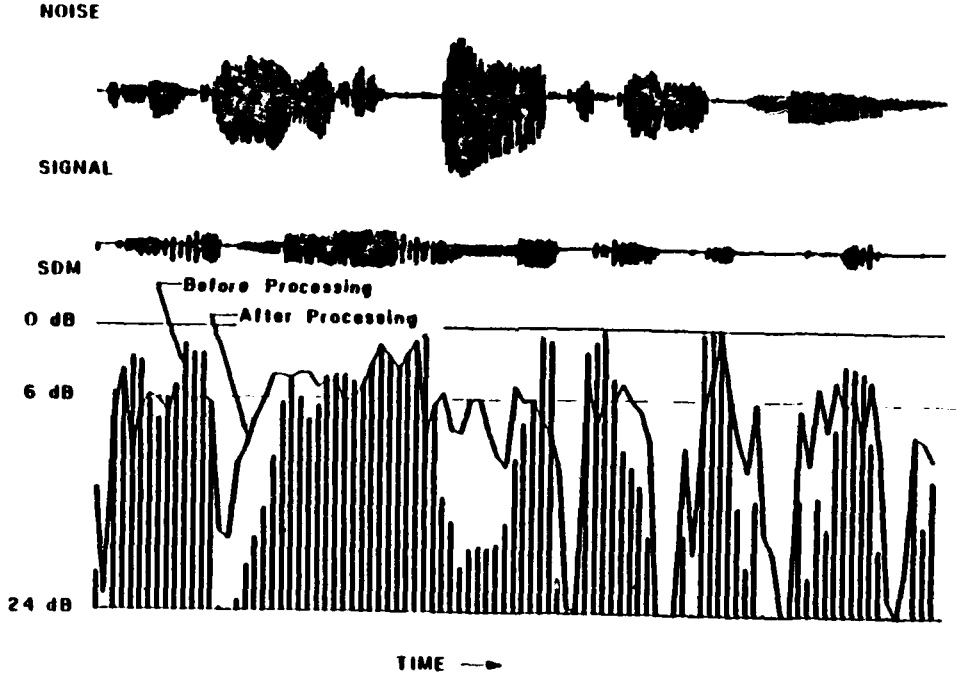


Fig. 2-3: Spectral Distortion Measure vs. Time
[SDM before (bars) and after (curve) processing]

speech signal as well as where the processing degrades the match. Such information has been found to be useful during algorithm development.

An overall performance measure of the processing algorithm can also be computed by averaging the short term SDM's over the length of the utterance:

$$SDM = \frac{1}{M} \sum_{m=1}^{M} SDM_{cb}(m) \qquad (2\text{-}8)$$

where

$SDM_{cb}(m)$ = short time SDM from equation (2-7) for m-th time interval (calculated every 20 msec)

The relation between SDM and SNR for ten unprocessed co-channel speech samples summed at various SNR's is shown in Fig. 2-4. Each point in this figure represents the SDM and SNR values (calculated from equation (2-8) and a simple energy ratio, respectively) for one co-channel sample consisting of desired and interfering speech of about 2 seconds duration. The spread of each sample group around the input SNR's (e.g. -6 dB, -9 dB) is a result of the pause removal in equation (2-1), which is not included in the simple energy ratio SNR (the abscissa of the figure).

It can be seen in Fig. 2-4 that SDM and SNR are highly correlated, which then implies that the SDM is correlated with the intelligibility of unprocessed co-channel speech.

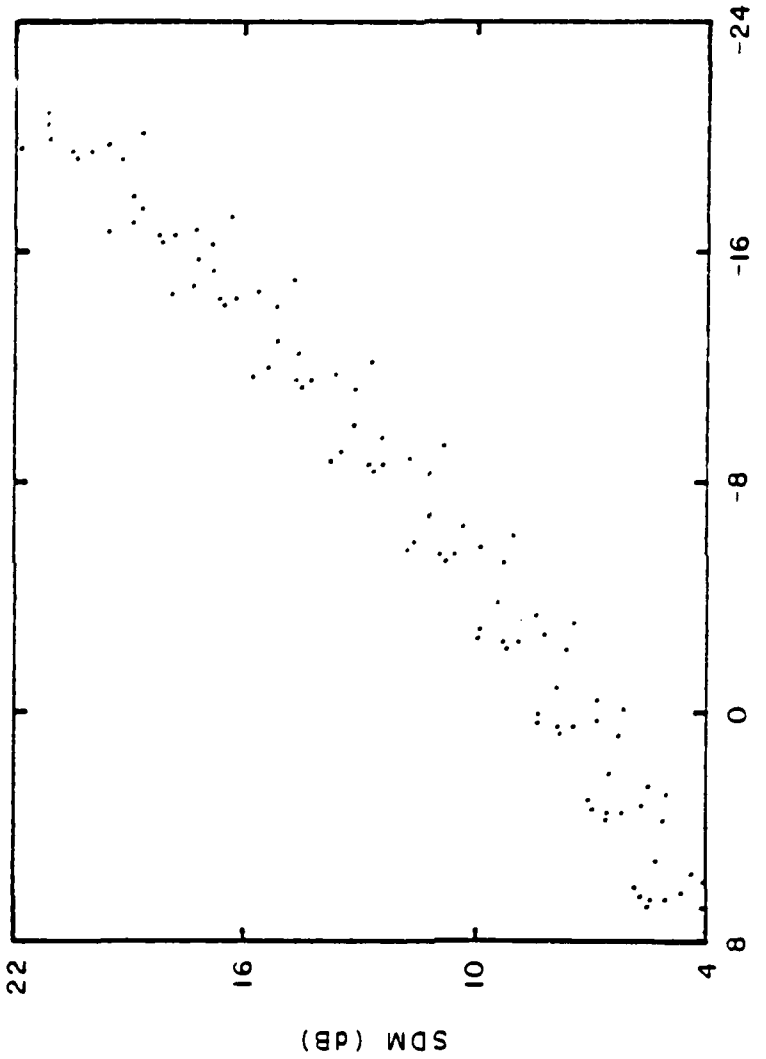


Fig. 2-4: SNR vs. SDM for Ten Unprocessed Co-Channel Speech Samples (input SNR varies from +6 to -20 dB)

Since the SDM does not require an estimate of the noise left after co-channel separation processing, it is also applicable to processed co-channel speech. Thus, SDM is a useful measure for estimating the intelligibility improvements obtained from the algorithms studied in this work.

3.0 SIGNAL HARMONIC EXTRACTION

As mentioned in chapter one, a number of speech separation techniques have been developed and tested in the past few years. This chapter presents the development and testing of a new extraction approach which incorporates the following features:

1. Signal pre-whitening with inverse filtering
2. Spectral magnitude harmonic sampling
3. Harmonic synthesis

Section 3.1 discusses the proposed extraction system and describes in detail its most important components. To evaluate this approach, a limited size intelligibility test was conducted and the results are presented in section 3.2. Careful analysis of these results, as discussed in section 3.3, provides new insights and directions for the speech separation problem that are applied in subsequent chapters.

3.1 A Pitch-Based Signal Extraction System

A signal in additive noise can be enhanced by either extracting the signal or suppressing the interference based on some consistent differences between the signal and noise characteristics. When the interference and signal are both speech, it is not possible to apply conventional filtering techniques because their long-term spectral characteristics are similar. Furthermore, since the short-term spectral

characteristics are most important when dealing with speech signals (see e.g. [Flanagan 1972, Rabiner and Schafer 1978]), the enhancement technique must make use of short-term differences.

One obvious short-term characteristic that can be exploited is the pitch contour from voiced speech. It can generally be assumed that the pitch contours of the desired and interfering speech are sufficiently separated so that the different pitch frequency harmonics are resolvable with short-term spectral analysis. This is illustrated in Fig. 3-1 which shows short-term spectra from two different speakers' voiced utterances. Note that sampling at speaker one's pitch harmonics generally misses the spectral peaks of the second speaker. A second important assumption of this approach is that sections where the pitches do overlap are short enough that the information carried in such segments can be deduced from neighboring segments based on syntax and semantics.

A total system approach which uses short-term spectral analysis of the signal pitch harmonics is shown in Fig. 3-2. The signal is first processed by a linear prediction coding (LPC) analysis and a pitch and voicing detection algorithm. It is then pre-whitened with the LPC inverse filter A(z). The unvoiced signal is replaced by white noise scaled by an estimated gain parameter. The voiced signal is processed

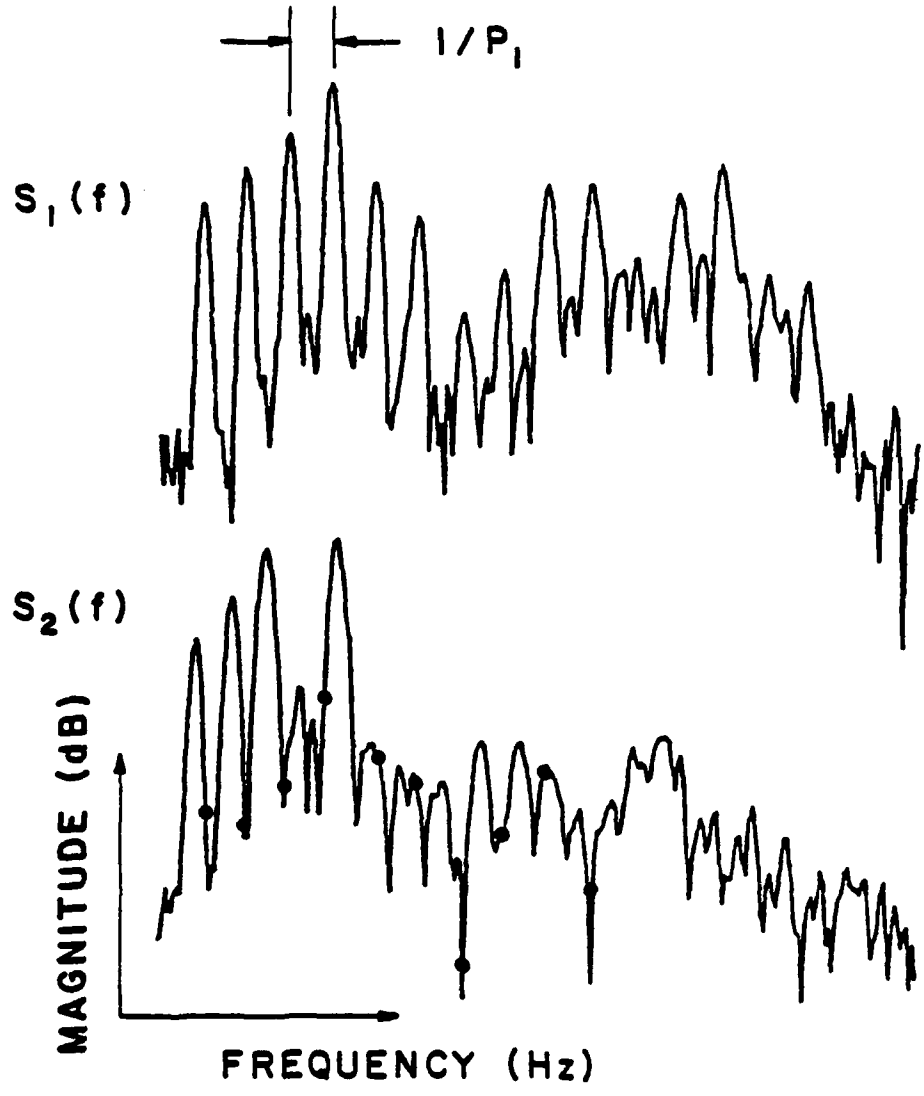


Fig. 3-1: Spectra of Voiced Speech from Two Speakers
(• indicate $|S_2(f)|$ samples spaced $1/P_1$ apart)

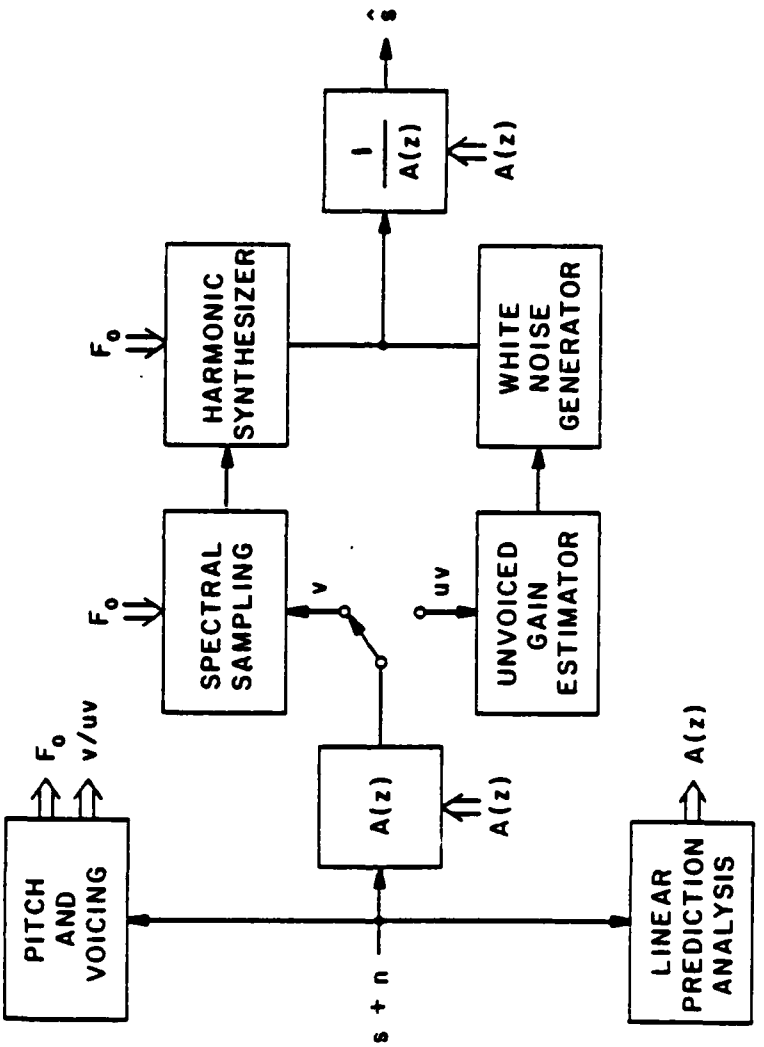


Fig. 3-2: Speech Signal Extraction System

with pitch harmonic sampling and synthesis algorithms. Both signals are then filtered by the all-pole filter 1/A(z).

There are three sets of problems that must be addressed in developing and testing the system of Fig. 3-2: (i) pitch and voicing detection on two-speaker speech, (ii) estimation of unvoiced speech level for the desired speaker, and (iii) harmonic sampling and synthesis of voiced speech. Although the first two problems are very important for the success of the system, the key to the system is the validity of the harmonic sampling and synthesis procedure. Therefore, in the experimentation discussed here, the first two problems are circumvented by using pitch and gain parameters estimated from speech free of interference. The details of the harmonic processing of the voiced speech are discussed in the next two subsections.

3.1.1 Spectral Pre-whitening and Sampling

Fig. 3-3 schematically illustrates a speech "analysis and synthesis" model where the inverse filter A(z) is calculated using LPC analysis [Markel and Gray 1976]. As can be seen, these models separate the input speech signal (represented by its z-transform S(z)) into what are referred to as its spectral envelope, A(z), and excitation (or residual), E(z), components.

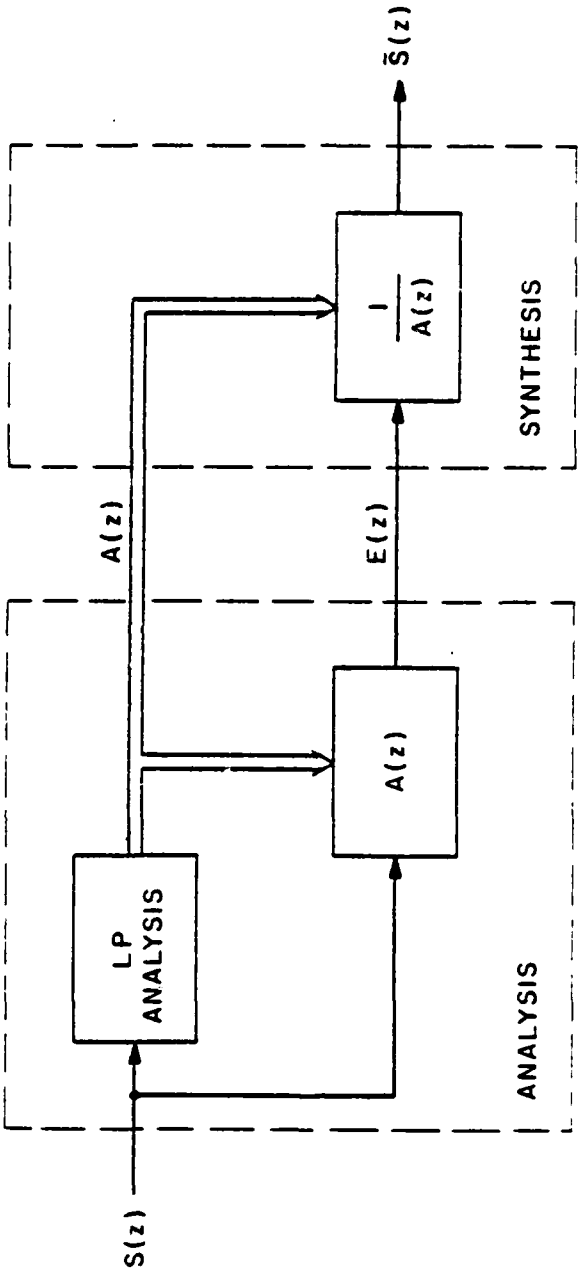


Fig. 3-3: Analysis and Synthesis Models of Speech

To evaluate the relative importance of the excitation and spectral envelope information in speech separation, two simple tests were run (these tests were originally proposed and reported by Juang [1981]). The corrupted signal s+n (desired speech plus interfering speech) and the clear speech s are deconvolved into an envelope model and an excitation signal by LPC analysis. Two output signals are then generated by driving each LPC synthesis filter with the other excitation signal. The output $\hat{s}_1$ is produced with excitation from s+n and the spectral envelope from s. The output $\hat{s}_2$ is produced with the spectral envelope from s+n and the excitation from s. The construction of $\hat{s}_1$ and $\hat{s}_2$ is illustrated in Fig. 3-4.

Informal listening tests were conducted to compare $\hat{s}_1$ and $\hat{s}_2$ for several different speech samples. Both outputs were found to sound much better than the unprocessed s+n signal. The result that the $\hat{s}_1$ output is intelligible is expected because exciting the desired speech envelope with only random noise is known to produce "whispered" but intelligible speech. What is significant, however, is that $\hat{s}_2$ actually sounds better than $\hat{s}_1$. This result suggests that harmonic processing to extract the desired speaker's residual signal may lead to better speech enhancement. Accordingly, as indicated in Fig. 3-2, LPC pre-whitening is performed before spectral sampling and harmonic synthesis,

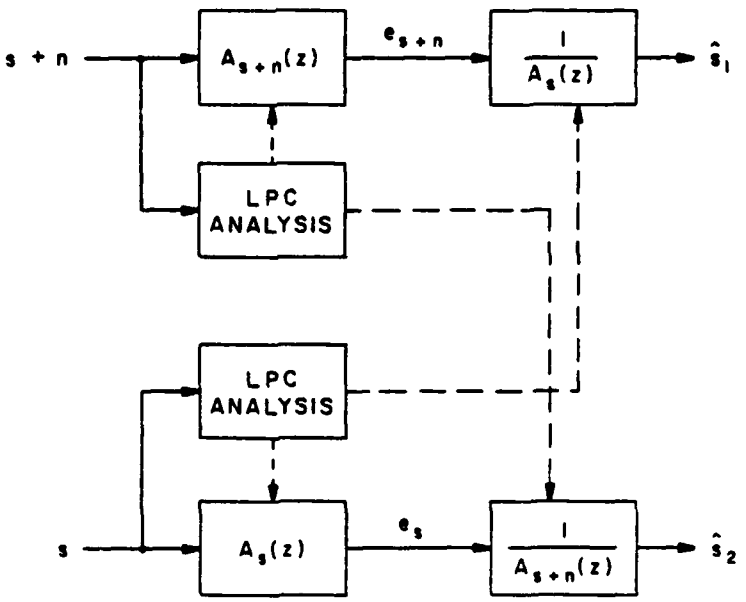


Fig. 3-4: Tests of Excitation and Envelope Parameters

and the spectral envelope filter is applied after the harmonic synthesizer.

Based on the assumption that the pitch frequencies of two unrelated voice signals (or residuals) generally do not overlap, the speech energy for a particular speaker would be concentrated at his/her harmonic frequencies. If the spectrum is sampled at the desired speaker's pitch harmonics, most of the energy of the spectrum samples would correspond to that speaker's voice. After obtaining the harmonic amplitudes, the desired time domain waveform is reproduced with the harmonic synthesis algorithm.

3.1.2 Harmonic Synthesis

The harmonic synthesis technique as described here was originally proposed by Markel and Gray [1978] as a possible solution to the problems of LPC synthesis at high pitch frequencies. In speech enhancement, this algorithm is useful since it avoids the problems with phase estimation from the noisy speech spectrum by generating a smoothed phase function from interpolated pitch values. This phase-generating feature, and the rest of the algorithm as developed by Markel and Gray, are described below.

Given that the harmonic amplitudes are known, a speech signal can be synthesized with a cosine series expansion:

$$s(n) = G \sum_{m=1}^{L} C_m \cos(m\theta_n + \phi_m) \qquad (3\text{-}1)$$

where

n = time index
C_m = spectral amplitude of the m-th pitch harmonic
ϕ_m = initial phase constants
G = gain
L = integer$[F_s/2F_o]$
(F_s = sample rate and F_o = pitch frequency)
θ_n = instantaneous phase for the first harmonic.

The initial phases at each harmonic ϕ_m, can be calculated from the speech spectrum. However, informal listening found that using all zero values for these phase constants gives the same degree of naturalness in the synthesis. The parameters C_m, G, and L are updated once for each N-point frame. In our experiments, the frame length is 20 msec. Assuming F_o is also updated once per frame (at n=0 and n=N for the current and next frames), then the intermediate pitch values are approximated by linear interpolation:

$$g_n = (g_N - g_0) \frac{n}{N} + g_0 \qquad (3\text{-}2)$$

The term g_n above can be viewed as the "instantaneous" pitch normalized by F_o, so the phase θ_n is approximated by summing g_n:

$$\theta_n = \theta_{n-1} + 2\pi g_n \qquad (3\text{-}3)$$

Continuity between frames is insured by setting θ_0 for the current frame to be θ_N of the previous frame. The harmonic amplitudes C_m are obtained by sampling the FFT magnitude, and the gain term is approximated using the input speech energy R_0:

$$G \approx \left\{ 2 \frac{R_0}{N} \Big/ \sum_{m=1}^{L} C_m^2 \right\}^{\frac{1}{2}} \qquad (3\text{-}4)$$

The approximation in equation (3-4) is due to the fact that the energy matching of the input speech with the synthesis is based on a fixed frame length which may not coincide with an integral number of pitch periods. For an exact energy match, the cosine series of equation (3-1) should be squared and summed over each frame, but the approximation of equation (3-4) was found to be accurate enough.

The harmonic synthesizer bears resemblance to the phase vocoder [Flanagan and Golden 1966]. Both systems consist of a set of filterbanks (the cosine terms in the harmonic synthesizer) controlled by magnitude and phase estimates. It differs from the phase vocoder in that the filterbanks are situated at the pitch harmonics, which makes them time-varying. Also, the harmonic synthesizer generates its phase information from pitch values, whereas the phase vocoder estimates phase directly from the short-term spectra of the input speech.

As Markel and Gray [1978] have pointed out, harmonic synthesis can be efficiently implemented if table lookups are used for the cosine functions. Since no filtering operation is carried out, filter instability problems, as in linear prediction synthesis, are avoided. However, the harmonic synthesizer cannot be applied for nonperiodic signals; other techniques (such as standard LPC analysis/synthesis) must be used instead. Because of this limitation, alternate processing for the unvoiced desired speaker segments is used in the extraction system of Fig. 3-2.

Prior to being incorporated into the speech extraction system of Fig. 3-2, the harmonic synthesizer was tested on voiced speech without interference. The synthesis from this "clean" speech was then evaluated with informal listening by several researchers, and was found to be generally equivalent in intelligibility and quality to LPC synthesis.

3.1.3 Effects of Phase

As the preceding subsection discusses, no explicit phase measurement is required for the the harmonic synthesizer to generate reasonable quality speech. This synthesis of speech without the exact phase information can be viewed as another example of G.S. Ohm's "acoustic phase law" [Schroeder 1975], which states that "'aural perception depends only on the amplitude spectrum of a sound and is independent of the phase angles of the various frequency

components contained in the spectrum'." This law generally applies to "short-time spectra" (e.g. $\leq$ 50 msec). Although exceptions to this phase law have been demonstrated in various experiments [Milios and Oppenheim 1983, Cox and Robinson 1980], most of these involve non-speech stimuli such as tones or long term phases. The main effect of phase on speech appears to be the quality of the synthesized speech (see e.g. [Wong 1979]).

In summary, while short-term spectral phase does have perceivable effects on speech quality, its effect on intelligibility is generally second order compared to spectral magnitude. In this study on co-channel separation algorithms, intelligibility is the first priority, hence the proposed techniques will only consider spectral magnitude information.

3.2 Testing and Results

The system described in section 3.1 was tested on several speech samples with voice interference. Informal listening found the output to be significantly enhanced in quality. To verify these qualitative judgments, formal evaluation was conducted using a limited-size intelligibility test. The purpose of the test was to evaluate the prewhitening and spectral sampling/harmonic synthesis parts of the system shown in Fig. 3-2. Therefore, the pitch, voicing, and gain contours were extracted from clean speech (using

standard vocoder algorithms).

The general method of intelligibility testing has been discussed in detail in section 2.1; a few specifics are listed here. The test data consisted of phonetically balanced sentences from male speakers with close and separated pitch contours added at average SNR's of 0 and -6 dB (representative pitch contours from the three speakers are shown in Fig. 3-5). These test sentences were then processed to extract the desired voice. For each test condition (SNR and pitch contour separation), one or two listeners were presented with ten speech samples, five processed and five unprocessed. The first listening procedure discussed in section 2.1 is used (i.e. an intelligibility comparison test). The percentage of correct words transcribed from the desired speaker were then compared for the processed versus the unprocessed data.

Single listener test scores are shown in Table 3-1. As might be expected, intelligibility is lower for the close pitch case and the lower SNR (-6 dB). The most significant result is that intelligibility scores are consistently lower for the processed speech. Although the test is limited in scale, the large intelligibility differences and the close correlation of these results with those of another study on a similar system [Perlmutter et al. 1977] suggest that more extensive testing is unnecessary. Given that further degra-

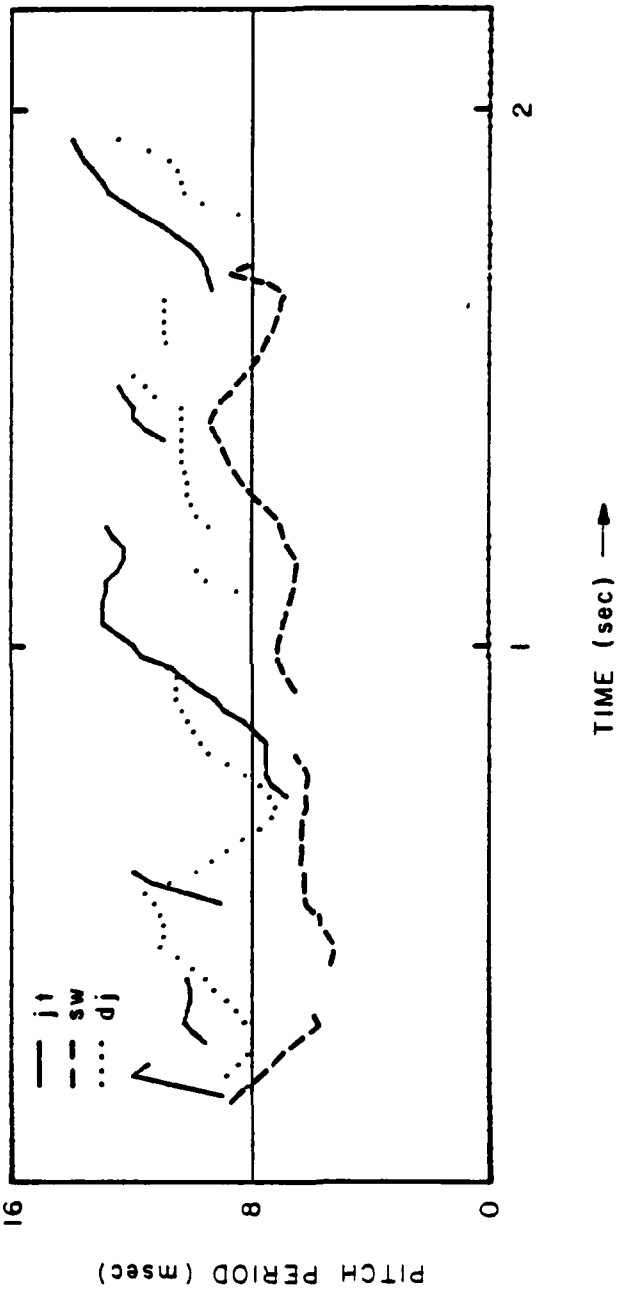


Fig. 3-5: Pitch Contours for the Three Speech Subjects

dation will be introduced due to pitch and gain estimation problems for corrupted speech, it is conclusive from these tests that the harmonic synthesis approach will not lead to intelligibility enhancement.

"Close Pitch" Speakers

	Unprocessed	Processed
-6 dB SNR	35.9	27.6
0 dB SNR	88.7	62.1

"Separated Pitch" Speakers

	Unprocessed	Processed
-6 dB SNR	75.2	43.3
0 dB SNR	87.3	67.6

Table 3-1: Intelligibility Scores
(% Correct Words)

3.3 Conclusions and New Directions

The lack of intelligibility improvement indicated by the testing was unexpected since informal listening had clearly found enhancement in the quality of the desired speech. The reason is that while processing does reduce the interference power, the desired speech also undergoes a considerable distortion in the synthesis process. The informal

listening subjects, who were already familiar with the test material, probably matched words to sounds, giving a false impression of intelligibility improvement. Thus the importance of carefully designed listening experiments cannot be overemphasized.

For voice interference, it has been shown here and in other work [Perlmutter et al. 1977] that speech above 0 dB average SNR is usually intelligible, but it degrades rapidly below 0 dB and is nearly unintelligible below -6 dB for "close" pitch cases. For "separated" pitch cases the desired speaker remains fairly intelligible down to even lower SNR values. Close examination of the test results presented in section 3.2 also finds 0 dB to be a significant intelligibility threshold for frame-by-frame "instantaneous" SNR, as illustrated in Fig. 3-6, which shows a typical transcription against the instantaneous SNR contour. Even though the average SNR is -6 dB for this case, there are short segments over which the instantaneous SNR is well above 0 dB, such as during speech peaks or noise pauses. Three of these segments with SNR > 0 dB coincide with the desired speaker's words "dull and tired" and were correctly transcribed. Similar correlations between such segments (i.e. with instantaneous SNR > 0 dB) and correct word transcriptions are found throughout the listening results for the unprocessed co-channel data. However, the same segments

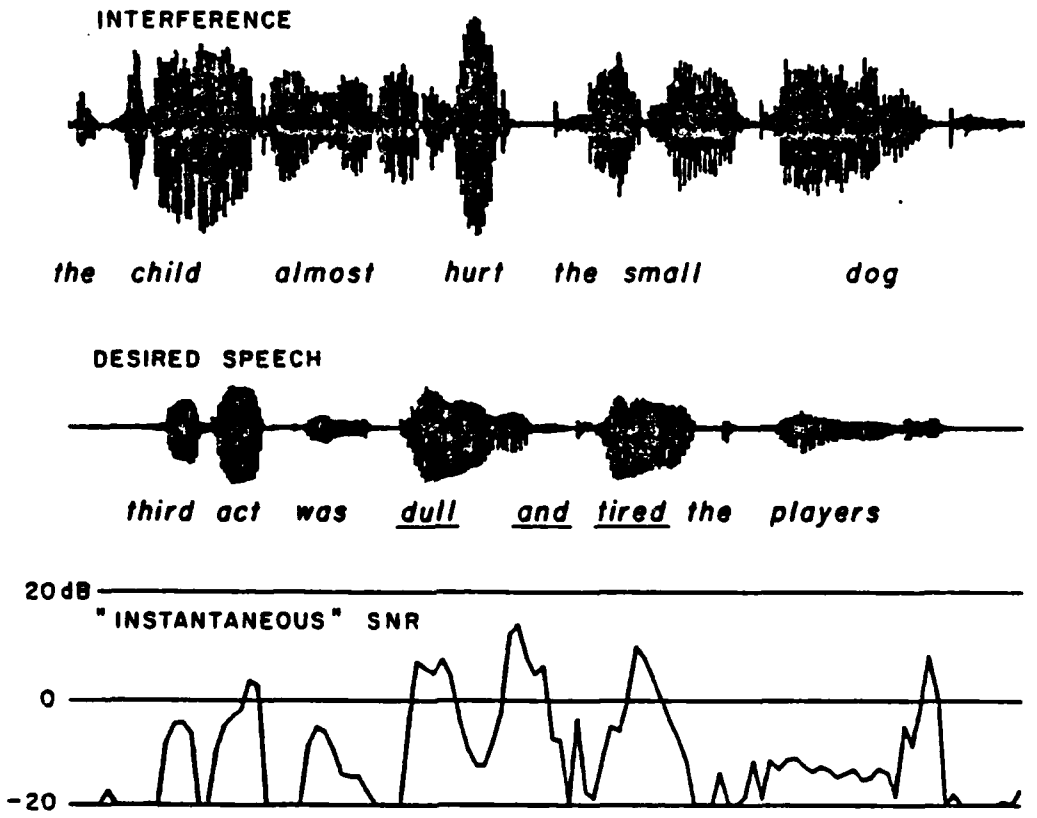


Fig. 3-6: Correlation of Correct Word Transcriptions (underlined) with Instantaneous SNR

are much less intelligible after harmonic processing.

A significant conclusion drawn from close examination of the test results is that when long term SNR exceeds 0 dB, it is best not to process the speech at all. Since co-channel-interfered speech with SNR's above 0 dB has been shown to be generally intelligible, the cases with the most potential for intelligibility improvement are those with average SNR's less than 0 dB.

Even for average SNR < 0 dB, the processing should be limited to segments where instantaneous SNR is under 0 dB. If such an SNR estimate could be obtained it would be very useful for switching the enhancement processing on and off so that only the lower SNR segments would be processed. This would avoid distorting the parts of the desired speech that are already intelligible. The importance of this control of the enhancement processing by the frame-to-frame characteristics of the input data (such as SNR) has also been suggested recently by Boll and Wahlford [1983], who proposed an "event driven speech enhancement" concept. Further study on this approach is highly recommended.

The new focus on negative SNR's in turn leads to the change in emphasis from signal extraction to noise suppression. That is, for negative average SNR, the interference is generally stronger, so its parameters, such as pitch, are more readily extractable. So the goal should be to extract

the interference signal parameters, such as pitch and harmonic amplitudes, which are more readily estimated, and use these parameters to remove the interference. Another important advantage of the interference removal approach is that it will generally leave the desired speech signal intact. It is very likely that the main reason the signal extraction technique of section 3.1 leads to degradation is that the desired speech signal has to be synthesized. Even without interference, the synthesized speech is noticeably degraded.

In summary, the new directions suggested by the results presented in this chapter are:

1. For average SNR > 0 dB, generally no processing is needed for all speech. Hence research should focus on average SNR < 0 dB cases.

2. The enhancement processing is generally needed only for speech segments with "instantaneous" SNR < 0 dB.

3. For negative SNR cases, interference suppression techniques should be applied instead of signal extraction.

4.0 CO-CHANNEL INTERFERENCE SUPPRESSION ALGORITHMS

Based on the results presented in chapter three, noise suppression algorithms for processing co-channel voice data with negative-decibel SNR were developed. The noise suppression algorithms developed in this chapter consist of two distinct components: the co-channel interference estimator and an algorithm that removes the estimated interference. The interference removal technique developed is the same for all the suppression algorithms, and is based on the spectral subtraction method. Accordingly, the first section of this chapter discusses the development of this spectral subtraction algorithm for co-channel interference removal. Sections 4.2 and 4.3 then discuss the development of several co-channel interference estimation approaches. Comparisons between the algorithms using spectral distortion measures and informal listening are presented in section 4.4.

4.1 Spectral Subtraction Concepts

4.1.1 Background

There has been much research on the use of spectral subtraction for enhancing noisy speech since its proposal by Weiss et al. [1974]. This technique has mainly been used for removing wideband noise from speech. Although no intelligibility improvement has been achieved for wideband noise, research is continuing on possible improvements to the

method [Nawab 1981, Hoy 1983]. This interest is probably due to the fact that spectral subtraction can improve the perceived quality of noisy speech, and it has demonstrated small gains in intelligibility when used as a preprocessor for LPC systems [Boll 1979].

The basic assumption of spectral subtraction, as it has been used for wideband noise reduction, is that noise and speech are uncorrelated processes. The noise power spectral density (PSD) is first estimated from the segments where there is no speech. Then the short-term energy spectrum of the desired speech is estimated by subtracting the (properly scaled) noise PSD from the short-term energy spectrum of the unprocessed noisy speech. These computations involve only the spectral energy because human perception is relatively insensitive to phase in the short-term spectra (as discussed in section 3.1.3). The final step consists of resynthesizing the desired speech waveform from the processed short-term magnitude spectra (the square-root of the estimated energy spectra) and the unprocessed phase.

These steps are illustrated by the diagram in Fig. 4-1, where $|\hat{N}|^2$ denotes the noise PSD estimated from the non-speech segments (as determined by the speech activity detector). The overlap-add (OLA) algorithm [Allen 1977, 1982] performs the post-subtraction inverse fast Fourier transform (IFFT) and smoothes over discontinuities at frame boundaries

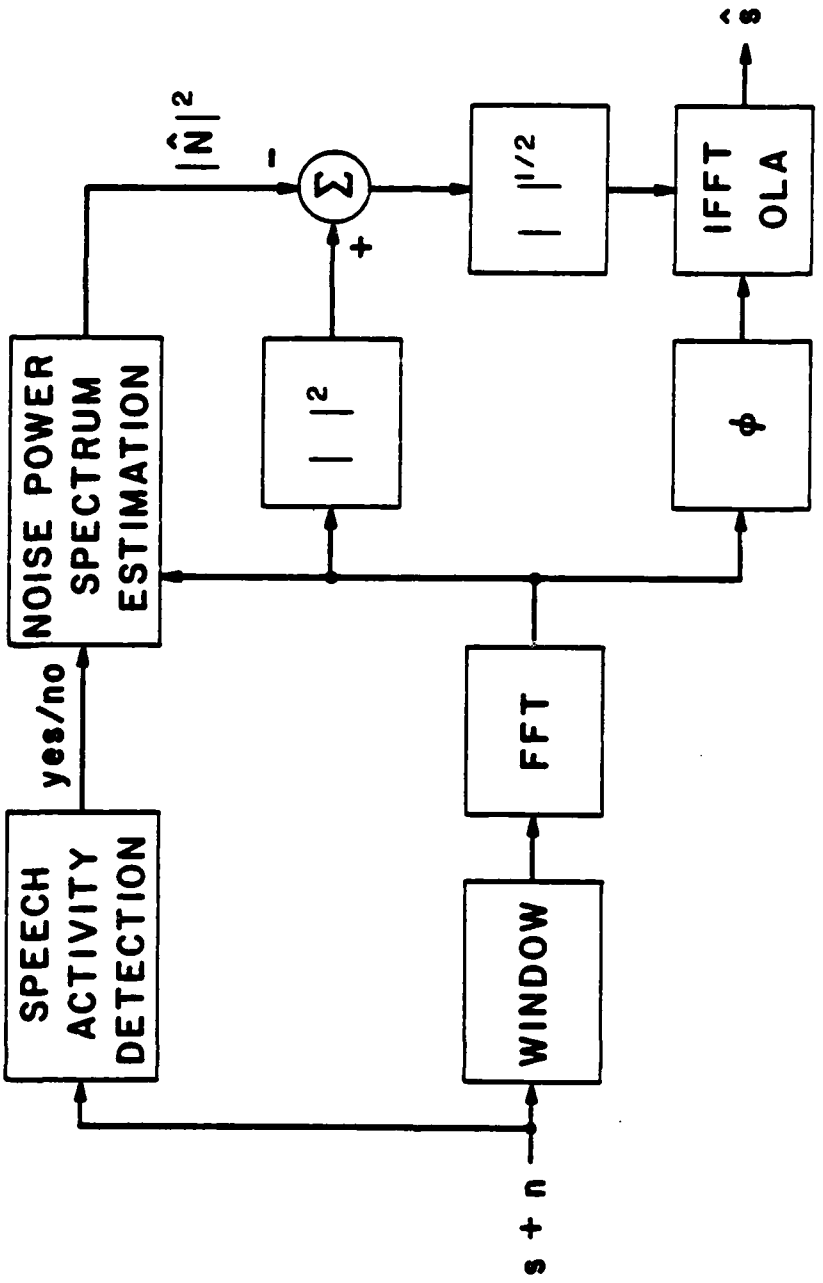


Fig. 4-1: Power Spectral Subtraction (for suppressing wideband random noise)

(heard as a continuous "buzz" at the frame frequency if OLA is not applied).

The "power spectral subtraction" technique discussed above may be generalized by raising the magnitude spectra to an arbitrary power, a, before subtraction and taking the "1/a"th root of the difference. The input to the OLA processing is then given by:

$$\hat{S}_w(f) = [|S_w(f)+N_w(f)|^a - |\hat{N}_w(f)|^a]^{1/a} \cdot e^{j\phi(f)} \qquad (4\text{-}1)$$

where:

a = exponent parameter
$\hat{S}_w(f)$ = estimated short-term spectrum of windowed desired speech [output signal is obtained by OLA processing of $S_w(f)$]
$\phi(f)$ = phase of windowed "noisy" speech, s_w+n_w

In the above, $S_w(f)$, $N_w(f)$ and $\hat{N}_w(f)$ represent the spectra of the windowed speech, noise, and estimated noise, respectively. Power spectral subtraction is implemented by setting a=2 in equation (4-1).

Note that if the estimated noise magnitude spectrum becomes larger than the magnitude spectrum of the windowed "noisy" speech at any frequency, it is possible to obtain a non-positive spectral difference in equation (4-1). Since the "1/a"th root of this spectral difference is interpreted as the magnitude of $\hat{S}_w(f)$, this situation must be avoided.

One solution is to set $\hat{S}_w(f)$ to zero for any differences less than zero, and this approach will be applied in this study (for simplicity this difference limiting will not be explicitly shown).

A formulation of spectral subtraction technique in terms of linear filtering (due to Paul [1979, 1981]) provides some interesting interpretations of the technique's operation. In his work on a robust vocoder algorithm, Paul shows that if the input to the spectral subtraction is the sum of the windowed signal and noise spectra,

$$X_w(f) = S_w(f)+N_w(f) \tag{4-2}$$

then the magnitude at the subtraction output, $\hat{S}_w(f)$, can be written as the product of the magnitude of this input with a filter magnitude function $|H(f)|$:

$$|\hat{S}_w(f)| = |H(f)|\ |X_w(f)| \tag{4-3}$$

where

$$|H(f)| = \left\{1- \frac{1}{[R(f)]^k}\right\}^{1/m} \tag{4-4}$$

$$R(f) = \frac{|X_w(f)|}{|\hat{N}_w(f)|}$$

By setting k=m=a, the above reduces to:

$$\hat{S}_w(f) = \left\{|S_w(f)+N_w(f)|^a - |\hat{N}_w(f)|^a\right\}^{1/a} \tag{4-5}$$

This is equivalent to the spectral difference term in the general spectral subtraction equation (4-1).

The term R(f) in equation (4-4) is a frequency dependent "signal plus noise to noise" ratio. Thus, it is apparent that the "filtering" indicated in equation (4-3) passes those spectral segments where this ratio is high (i.e. strong signal and weak noise), while suppressing segments where it is low (i.e. weak signal and strong noise). In fact, minimum mean square error filtering is obtained for stationary and uncorrelated signal and noise if k=2 and m=1. Then equation (4-3) reduces to an estimate of the noncasual Wiener filter:

$$|H(f)| = \frac{|S_w(f)|^2}{|S_w(f)|^2 + |\hat{N}_w(f)|^2} \qquad (4\text{-}6)$$

4.1.2 Analysis of Exponent Parameter

Referring again to the general equation for spectral subtraction given in (4-1), the influence of the exponent parameter "a" on the results should be analyzed to determine the proper value for implementation. In previous research [Lim 1978, Berouti et al. 1979, and Paul 1979, 1981] different values of this parameter have been tried with varying degrees of success for wideband noise situations. For example, Lim tried 2, 1, .5, and .25 for "a" and found that for constant SNR the intelligibility of the recovered speech

decreased monotonically with the exponent parameter value. While the results of these earlier researchers may not be directly applicable to the co-channel interference case, they do suggest that the exponent parameter requires careful study. In this section a derivation is presented which considers the effects of the exponent parameter for the low SNR case. The results of this analysis suggest that magnitude difference may be preferable to other types of subtraction in this case.

The exponent parameter affects only the magnitude term in equation (4-1), so denoting this difference as D(f), then:

$$D(f) = \left\{ |S_w(f)+N_w(f)|^a - |\hat{N}_w(f)|^a \right\}^{1/a} \qquad (4\text{-}7)$$

If the difference in phase between $S_w(f)$ and $N_w(f)$ is defined as $\theta(f)$, then the magnitude of the sum can be expanded (the "(f)" are dropped here to simplify the notation):

$$D(f) = \left\{ (|N_w|^2 + |S_w|^2 + 2|N_w||S_w|\cos\theta)^{a/2} - |\hat{N}_w|^a \right\}^{1/a} \qquad (4\text{-}8)$$

If it is assumed the noise doesn't go to zero (i.e. $|N_w|>0$), it can be factored out:

$$D(f) = \left\{ |N_w|^a \left\{ 1 + \left(\frac{|S_w|}{|N_w|} \right)^2 + 2 \left(\frac{|S_w|}{|N_w|} \right) \cos\theta \right\}^{a/2} - |\hat{N}_w|^a \right\}^{1/a} \qquad (4\text{-}9)$$

Equation (4-9) illustrates the dependency of the processing on the SNR. Now assume that SNR << 0 dB. Making a second assumption that θ is not close to $\pm\frac{\pi}{2}$, the squared SNR term in equation (4-9) becomes insignificant compared to the linear SNR term:

$$\left(\frac{|S_w|}{|N_w|} \right)^2 << 2 \frac{|S_w|}{|N_w|} \cos\theta \qquad \text{(for SNR << 0 dB)} \qquad (4\text{-}10)$$

Dropping the squared SNR from equation (4-9) and using the first two terms in the Taylor expansion (again assuming SNR << 0 dB) yields:

$$D(f) \approx |N_w|^a + a|N_w|^a \frac{|S_w|}{|N_w|} \cos\theta - |\hat{N}_w|^a \qquad (4\text{-}11)$$

If a good estimate of the noise spectrum is available, then $|\hat{N}_w| \approx |N_w|$, and:

$$D(f) \approx a|N_w|^{a-1}|S_w|\cos\theta \qquad (4\text{-}12)$$

Consider the effect of selecting several different values of the exponent parameter a:

$$a = 2 \quad \text{(power diffs):} \qquad D(f) \approx 2|N_w||S_w|\cos\theta \qquad (4\text{-}13)$$

a = 1 (mag. diffs): $D(f) \approx |S_w|\cos\theta$ (4-14)

a = 0.5 (sqrt. diffs): $D(f) \approx 0.5|S_w|\cos\theta/\text{sqrt}|N_w|$ (4-15)

For all cases except the a=1 case the spectral difference is multiplied by $|N_w|$, the magnitude of the noise spectrum. For broad-band noise this multiplicative factor is not an important factor because $|N_w|$ is nearly constant for all frequencies. However, when the noise is speech (which usually does not have a "flat" spectrum) the multiplication factor $|N_w|$ can result in considerable spectral distortion. The phase difference between the signal and noise also affects D(f), through the $\cos\theta(f)$ term, but this term is present for all values of the exponent parameter a. In our listening tests, which will be described in the next section, the $\cos\theta(f)$ term by itself (i.e. in the a=1 case) does not seriously affect the intelligibility of the inverse transform of D(f) (which gives the spectral subtraction output signal).

4.1.3 Spectral Subtraction Implementation and Testing

Before implementing a noise suppression system based on spectral subtraction, it is necessary to determine whether spectral subtraction is a valid approach for suppressing co-channel speech interference. The experiment presented below considers the case where the interference magnitude spectrum is available. The purpose of this experiment is to

first validate the use of "noisy" phase in the synthesis process of the spectral subtraction algorithm for co-channel speech. Secondly, this experiment compares the performance of spectral subtraction for several values of the exponent parameter "a".

The algorithm used for this evaluation is illustrated in Figure 4-2. It is derived from the PSD subtraction shown in Fig. 4-1. In this case the noise spectrum estimate is calculated from $\hat{n}(t)$ (an estimated noise signal) and not from the silence segments as indicated in Fig. 4-1. A continuous noise estimate is required here because the noise signal is not stationary. To verify the analysis of section 4.1.2, where it is shown that a=1 gives the spectral difference with the fewest distortion factors, values of a = 1, 2, and 0.5 are used.

Consideration was originally given to alternative transforms instead of the FFT, as suggested by a number of recent studies. Petersen [1980] suggests that constant-Q transforms are more appropriate because of their closer modeling of auditory processes. McAulay and Malpass [1980] take a similar approach by using an "increasing-bandwidth-with-frequency" filterbank in their modified spectral subtraction algorithm. However, both of these papers are concerned with removing wideband noise and not with speech interference. The noise estimation and suppression

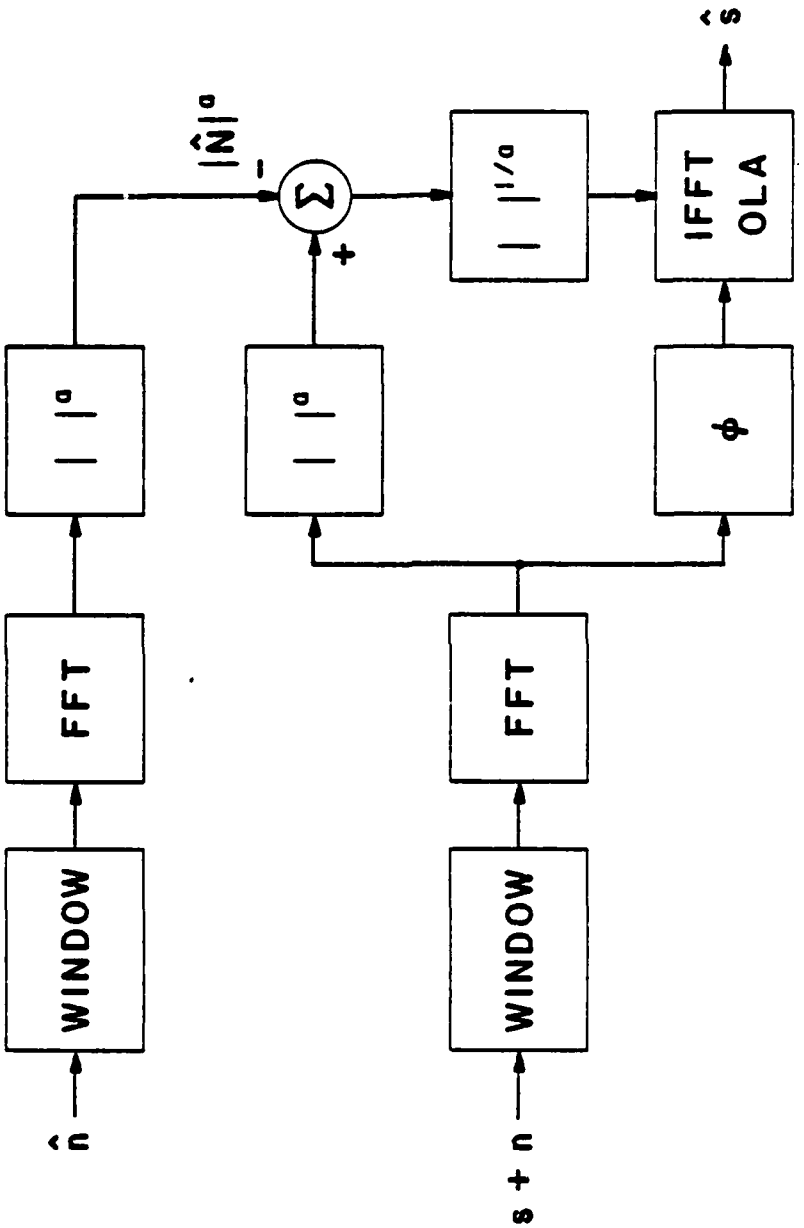


Fig. 4-2: Algorithm for Spectral Difference and Resynthesis Evaluation

approaches for speech interference rely on resolving the individual pitch harmonics of the interference. The resolution afforded by "constant-Q" transforms is not sufficient. Thus, standard FFT's are used for spectral estimation.

A Hamming window is applied to the input data because of its preferred tradeoff of bandwidth versus leakage suppression. This window is also compatible with the overlap-add processing used at the output [Allen 1977, 1982]. The mainlobe and first few sidelobes of the magnitude frequency response of a Hamming window to a sinusoid of frequency f_1 are indicated in Fig. 4-3. As can be seen, the mainlobe is 4/T Hz wide, where T is the window length in seconds, so the spectral resolution improves with increasing window lengths. Unfortunately, speech is not a stationary process, so the window has to be relatively short in order to capture enough time resolution. A reasonable compromise between minimum window length and spectral resolution is a 40 msec window.

At the system sampling rate of 10 kHz, a 40 msec Hamming window corresponds to 400 data samples, which require a 512-point FFT for the transform. With the 50% overlap used here for the overlap-add processing, the FFT's and spectral subtraction are done every 20 msecs. This gives a satisfactory degree of temporal resolution since vowel speech spectra are relatively invariant over a 20 msec interval.

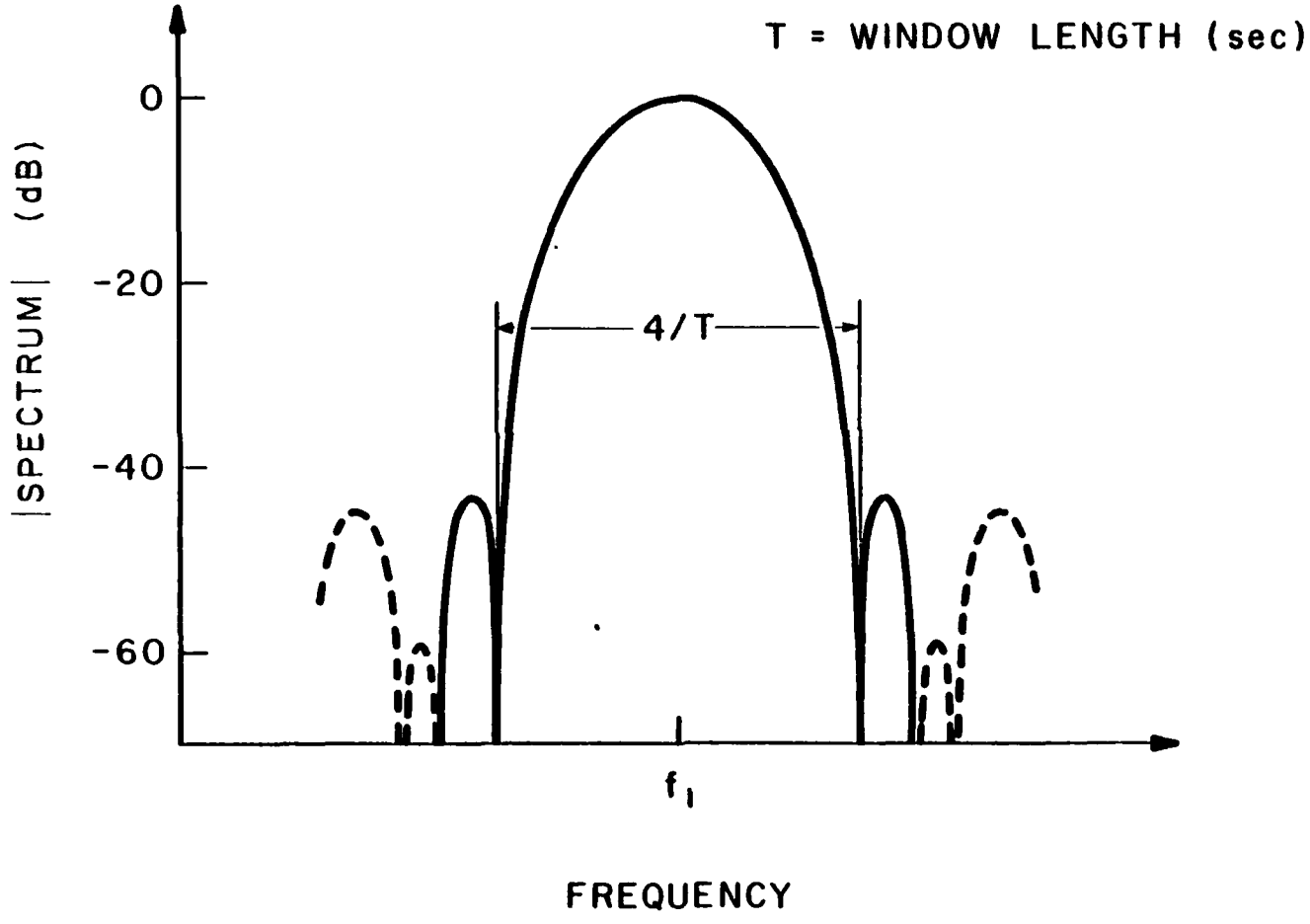


Fig. 4-3: Spectrum of Hamming-Windowed Sinusoid of Frequency f_1 (mainlobe and first sidelobes)

The system was first checked with several test signals; these consisted of speech with various additive tones and wideband noise. Then co-channel interference speech samples with SNR's ranging from -40 to -6 dB were processed. The outputs for numerous cases with the different a-parameters of 2, 1, and 0.5 were compared through informal listening and with the spectral distortion measures discussed in chapter two.

Typical results from the tests are given in Table 4-1 for an input SNR of -20 dB. The results show that the magnitude subtraction gives the lowest spectral distortion, with power subtraction a close second, and root magnitude showing the highest spectral distortions. Informal listening finds that very little interference is perceivable after spectral magnitude (a=1) subtraction. A moderate degree of distortion is heard, but intelligibility is not perceived to be affected. In contrast, the power (a=2) subtraction output contains a significant amount of residual interference and sounds less intelligible. The root magnitude (a=0.5) subtraction results also sound less intelligible than the magnitude data (however the root magnitude data appears to contain less residual interference than the power subtraction output). The speech quality is particularly poor over lower amplitude segments (such as voiceless consonants and ends of words). The root magnitude subtraction output also

SUBTRACTION TECHNIQUE	PROCESSED SAMPLE NUMBER						COMMENTS
	SDM in dB (18 dB limit SDM)						
	#3	#4	#5	#6	#7	#8	
Mag. sub [a=1]	4.9 (4.9)	5.0 (5.0)	4.6 (4.6)	4.8 (4.7)	5.1 (5.0)	5.2 (5.2)	Most intelligible sounding, very little background noise; some signal distortion
Power sub [a=2]	6.7 (6.3)	6.5 (6.0)	7.5 (7.0)	7.5 (6.8)	7.6 (6.6)	6.5 (6.0)	Much background noise; some signal distortion
$\sqrt{\text{mag sub}}$ [a=1/2]	21.2 (15.6)	19.7 (14.7)	20.7 (14.6)	20.2 (15.0)	20.8 (15.2)	20.4 (15.4)	Much distortion, particularly during lower amplitude segments of desired speech; "musical tone" noise

Test setup:
- -12 dB SNR
- interfering speaker is dj in samples 3, 4, and 5
- interfering speaker is jt in samples 6, 7, and 8

Table 4-1 Exponent Parameter Tests
(spectral distortion measures and informal listening)

contained "musical tones" type background noise that is well known in previous wideband noise spectral subtraction research (see e.g. [Berouti et al. 1979, Wong 1979]).

The -20 dB SNR tests discussed above illustrate the effect of the power parameter "a" on the output speech. The magnitude subtraction (a=1) is found to perform better than the other selections. The same result has been found to different degrees over a wide range of SNR values (i.e. -40 to -6 dB).

4.1.4 Discussion

The experiment presented in 4.1.3 shows that magnitude differencing is the preferred spectral subtraction technique for co-channel interference suppression. Experiments with spectral subtraction algorithms which use estimated interference spectra have confirmed this result. Details will be discussed in the following sections of this chapter.

More important than the selection of the difference power "a" discussed above, is the conclusion, derived from the experiments in section 4.1.3, that spectral subtraction successfully suppresses co-channel interference using only spectral magnitude information from the interference. The lack of accurate phase information in the resynthesis operation of spectral subtraction was initially thought to be a possible source of error. However, since the tests done

here show good intelligibility down to -40 dB SNR, the relative importance of phase is seen to be negligible.

Another point illustrated by this study is the effect of the cross-spectral magnitude term of the signal plus noise magnitude spectrum, $|S_w+N_w|$, rewritten below:

$$|S_w+N_w| = \left\{|S_w|^2 + |N_w|^2 + 2|S_w||N_w|\cos\theta\right\}^{1/2} \qquad (4\text{-}16)$$

where:

$$\theta = \text{phase}(S_w) - \text{phase}(N_w)$$

It was originally thought that the cross term (i.e. $2|S_w||N_w|\cos\theta$) was the source of error in the spectral difference calculation. However as the derivation of section 4.1.2 shows, if a good estimate of the noise spectral magnitude is available, then for SNR << 0 dB the desired signal magnitude spectrum is actually carried in the cross term.

4.2 Spectral Subtraction with Interference Synthesis

The preceding section investigated spectral subtraction for co-channel interference suppression assuming a good estimate of the interfering speech magnitude spectrum is available. The rest of this chapter considers the other half of the problem (i.e. estimating the co-channel

interference). Several interference estimation methods are developed and combined with spectral subtraction. In sections 4.2.1 and 4.2.2, two time-domain noise estimation techniques, LPC and harmonic synthesis, are presented.

All of the interference estimation techniques developed are pitch-based, so the same assumptions (infrequent overlap of desired and interfering talkers' pitch contours, etc.) made for the pitch-based extraction algorithm of chapter three are applicable. The primary difference is that the pitch-based processing is now used to estimate and suppress the noise. For the negative SNR conditions under consideration, the assumption that good pitch estimates are available is actually more reasonable (i.e. the pitch is now calculated for the interference which is the higher energy part of the co-channel signal).

Pitch-based interference estimation and suppression applies only to voiced segments of the interference, which are generally higher in energy than the unvoiced (non-harmonic) segments. Unvoiced interference segments are also difficult to estimate on a short-term basis because of their broadband noise character. Hence no attempt is made in this study to estimate and eliminate unvoiced interfering speech.

4.2.1 Spectral Sampling/Harmonic Synthesis (SS/HS)

The harmonic synthesis algorithm described in chapter three is used to obtain an interference estimate for spectral subtraction by spectral sampling at the interference pitch harmonics. A block diagram of this approach is shown in Fig. 4-4. The "spectral magnitude subtraction" component represents the spectral difference and resynthesis operations of Fig. 4-2, with a=1 for magnitude differences. The noise estimate, $\hat{n}$, for this subtraction comes from the harmonic synthesizer, which in turn uses the estimates of the noise energy and pitch harmonic amplitudes determined from the spectral sampling (R_0 and C_m of equations (3-4) and (3-1), respectively). Since the spectral sampling algorithm requires computation of the same windowed "s+n" FFT used in spectral magnitude subtraction, the FFT output is used for both operations.

The output of the system is switched between the spectral magnitude subtraction output and the original co-channel interfered speech, s+n. When the interference is unvoiced, s+n is simply passed through the system. Linear interpolation between the two switch positions is performed to reduce discontinuities caused by voicing changes. For example, when the interference changes from voiced to unvoiced speech, the output is interpolated over "M" data points around this transition using:

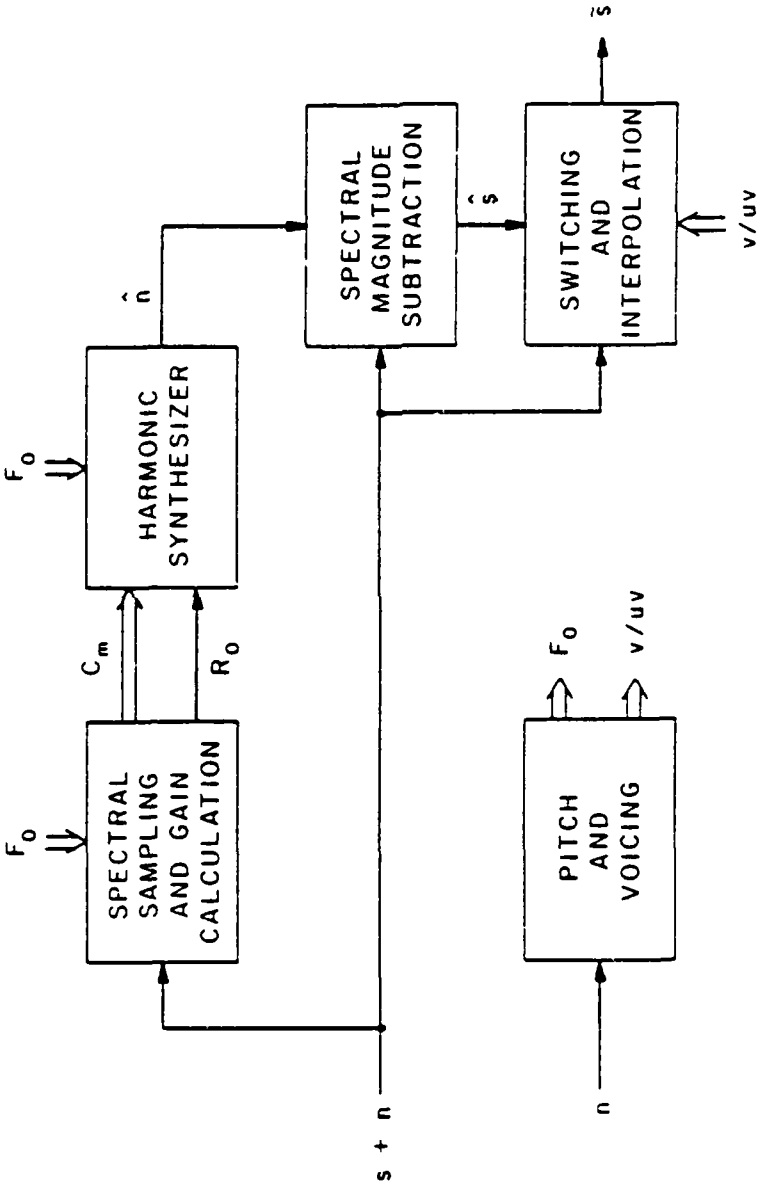


Fig. 4-4: Spectral Subtraction Using Noise Estimated from Spectral Sampling/Harmonic Synthesis

$$\tilde{s}(m) = \left(\frac{M-1-m}{M-1}\right)\hat{s}(m) + \left(\frac{m}{M-1}\right)(s(m)+n(m)) \qquad (4\text{-}17)$$

The length of the "frame" or "block" of data in the above interpolation, and for all operations of the noise estimation, is 20 msec (i.e. M=200); this "frame" length equals the interval between successive spectral subtraction operations, as detailed in 4.1.3.

An important parameter that can be varied in taking spectral differences, but has not yet been discussed, is the gain factor, g_s [Berouti et al. 1979, Wong 1979]. This can be included in the spectral difference of equation (4-1) as a multiplier of the estimated noise spectrum:

$$|\hat{S}_w| = \left\{ |S_w(f)+N_w(f)|^a - g_s|\hat{N}_w|^a \right\}^{1/a} \qquad (4\text{-}18)$$

To incorporate this parameter into the present system, its square (i.e. g_s^2) can be inserted as a multiplying factor of R_0 in Fig. 4-4. A value of one was assumed for g_s in section 4.1, but other values can be used if it becomes necessary to compensate for any consistent scale error in the level of the estimated noise spectrum.

A range of values was tried for g_s, but no value appeared to give significant improvement over the original value of g_s=1. It is interesting to note that if g_s is made too large (i.e. $g_s > 2$), "musical tone" noise is generated.

This is to be expected, since such "over subtraction" tends to leave isolated non-zero spectral components.

Spectral distortion measure (SDM) comparisons were done between power, magnitude, and root magnitude spectral subtraction with the harmonic synthesis noise estimation, and Table 4-2 summarizes the findings. The magnitude spectral subtraction (a=1) again yields the best overall results in the SDM. However, since the interference spectra used here are estimated, the distinctions between the three types of spectral subtraction are not as pronounced as in the exact noise spectral subtraction tests of Table 4-1. The magnitude and root magnitude subtraction SDM's are particularly close.

Informal listening comparisons were also conducted. The listening evaluation found that the magnitude subtraction cases sound better than the root magnitude data; both methods reduce the interference, but in the root magnitude cases the quality and gain characteristics of the desired speech are more distorted. Thus, magnitude spectral subtraction again appears to be a better choice than power or root magnitude subtraction.

Comparisons to the other two interference estimation algorithms will be discussed in section 4.4.

CASE	PROCESSED SAMPLE NUMBER						COMMENTS
	SDM in dB (18 dB limit SDM)						
	#3	#4	#5	#6	#7	#8	
Mag. sub [a=1]	9.5 (8.0)	9.7 (7.5)	11.0 (8.4)	10.3 (8.0)	11.7 (8.2)	9.2 (7.5)	Interference reduced; some distortion of desired speech
Power sub [a=2]	10.5 (8.2)	11.0 (8.1)	12.4 (8.7)	11.9 (8.6)	12.5 (8.5)	11.5 (8.3)	Interference more apparent than mag cases
$\sqrt{\text{mag sub}}$ [a=1/2]	10.1 (9.2)	9.9 (8.4)	11.0 (9.4)	9.5 (8.2)	11.5 (9.0)	8.6 (7.6)	Desired voice distorted by "echoing"; gain contours also significantly modified

Test setup:
- -12 dB SNR
- interfering speaker is dj in samples 3, 4, and 5
- interfering speaker is jt in samples 6, 7, and 8

Table 4-2 Exponent Parameter Tests for Spectral Subtraction with Harmonic Synthesis

4.2.2 LPC Noise Synthesis (LPCN)

The algorithm considered in this section is almost the same as the SS/HS algorithm just described. It is different only in that the interference is estimated by LPC analysis and synthesis. A block diagram of the system is shown in Fig. 4-5(a).

To evaluate the potential of this technique, LPC analysis/synthesis of the interference <u>alone</u> is first obtained as shown in Fig. 4-5(b). The "clean noise" LPC synthesis is used to suppress the interference spectrum by magnitude spectral subtraction. This experiment provides testing of spectral subtraction for noise estimates which approximate the noise spectrum in envelope characteristics, $\sigma/A_n(z)$, and pitch frequency spacing of the harmonics, F_o.

The test system of Fig. 4-5(b) resulted in a significant amount of interference suppression according to informal listening. However the amount of interference suppression is much less than the near perfect results of the "exact noise magnitude" tests described in section 4.1. The difference is a result of errors in LPC synthesis modeling of speech.

The next experiment determines whether adequate interference suppression can be obtained with LPC noise synthesis obtained by combining a F_o contour from "clean noise" (i.e. n) and an LPC spectral model derived from s+n

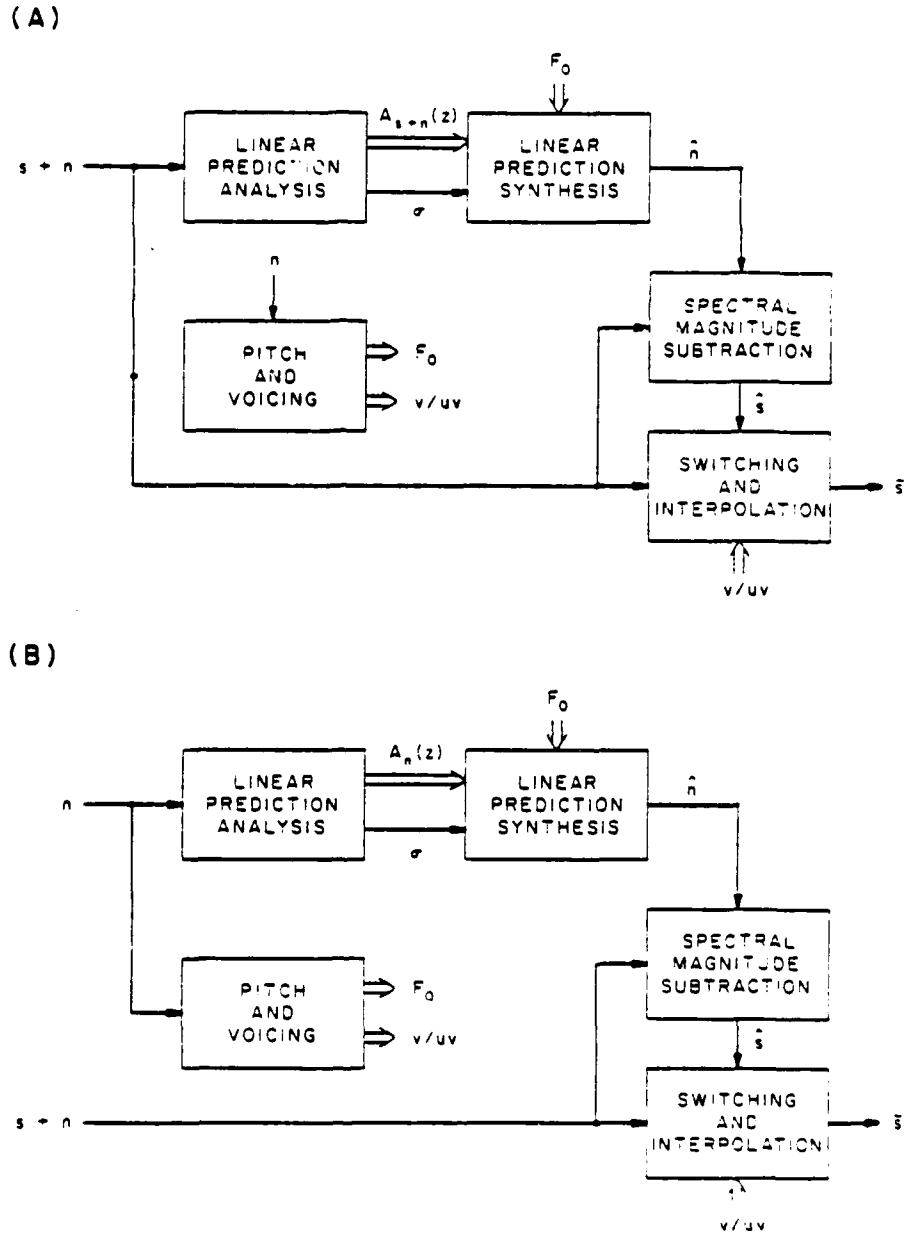


Fig. 4-5: Spectral Subtraction Using Estimated Noise from LPC Analysis/Synthesis
(A) LPC analysis/synthesis from "s+n"
(B) LPC analysis/synthesis from noise alone

(i.e. $A_{s+n}(z)$). It was originally expected, with the assumption of low SNR, that $A_{s+n}(z)$ would be sufficiently close to $A_n(z)$ that derivation of the LPC parameters from "s+n", as in Fig. 4-5(a), would yield similar results to Fig. 4.5(b). Also, the relative importance of the LPC residual signal with respect to the envelope, shown by the experiments discussed in section 3.1.1, suggests that if a residual signal obtained from the "clean noise" pitch were used to excite $A_{s+n}(z)$, a good noise estimate could be synthesized. Unfortunately, the results obtained using $A_{s+n}(z)$ were substantially worse than the results from using the LPC noise synthesis obtained from noise only.

Several modifications for improving the output quality were investigated. First, since the total squared error of the spectral modeling with LPC is known to decrease as the filter order M is increased, a range of filter orders up to M=24 was evaluated. As M approaches infinity, the model spectrum approaches the short-term magnitude spectrum of the input [Markel and Gray 1976]:

$$\sigma / |A_{s+n}(z)| \rightarrow |S(z)+N(z)|, \text{ as } M \rightarrow \infty \qquad (4\text{-}19)$$

$$\text{where } z = e^{j2\pi f/F_s}$$

So it was expected that for large M the LPC-synthesized noise technique will give similar results to the SS/HS considered in section 4.2.1 (as shown in section 4.4, the

spectral distortion computations for this technique do closely resemble those from SS/HS). However, varying M between 12 and 24 makes no significant difference in the results, so M=12 is used for the comparisons of section 4.4. Next, window overlap in the spectral subtraction was decreased from 20 msec to 10 msec in order to obtain better time resolution, but this also did not significantly affect the spectral distortion performance. Finally, the gain and exponent parameters (g_s and a in equation (4-18)) were varied. These tests again showed that the preferred parameter values are those initially used (a=1 and g_s=1).

Comparison to the other interference estimation techniques will be presented in section 4.4.

4.3 Harmonic Magnitude Suppression (HMS)

The basic premise of the HMS algorithm is that pitch harmonic spectral sampling can be used to estimate the noise magnitude spectrum for spectral subtraction, as done previously with the SS/HS technique of section 4.2.1. However, the HMS approach exploits several properties of the situation to obtain better estimates of the interfering speaker's magnitude spectrum:

1) Steady state voiced (periodic) segments of the speech interference can be expressed as a sum of harmonics. Thus, interference magnitude spectrum can be estimated from an approximation of a spectrum of win-

dowed sinusoids, with amplitudes determined from 2) below (this harmonic property is not accurate for voiced speech segments where the pitch is changing rapidly; however, in most cases the pitch is fairly constant over a short window).

2) The best estimate of the amplitude of each interference harmonic is obtained at the peak of the harmonics (i.e. at integer multiples of fundamental pitch frequency).

3) Pitch estimation errors of the voiced interference are generally small (a few Hz). An adaptive procedure using a minimum spectral difference power optimality criterion is developed to correct such errors.

Harmonic Sampling

Consider modeling a voiced interfering speech segment of constant pitch frequency F_o by a sum of cosines:

$$n(m) = \sum_{p=1}^{L} D_p \cos(mf_p + Y_p) \qquad (4\text{-}20)$$

where:

m = time index

D_p = spectral amplitude of p-th pitch harmonic

Y_p = phase of p-th pitch harmonic

L = integer $[F_s/2F_o]$ (F_s = sample rate)

$f_p = 2\pi pF_o/F_s$ (normalized pitch harmonic frequency)

To measure the spectral amplitude values, D_p, the signal is time limited with a finite length time window, w(m), and discrete Fourier transformation (DFT) is performed on the product (the "w" subscript on $N_w(k)$ indicates this

windowing):

$$N_w(k) = \sum_{m=0}^{M-1} n(m)w(m)e^{-jmk2\pi/M} \qquad (4\text{-}21)$$

Substituting the expansion for n(m) of equation (4-20) into (4-21), denoting the transform of w(m) by W, and using convolution in the frequency domain for the time domain product yields:

$$N_w(k) = \frac{1}{2}\sum_{p=1}^{L} D_p e^{j\Upsilon_p} W\left\{e^{j(\theta-f_p)}\right\} + \frac{1}{2}\sum_{p=1}^{L} D_p e^{-j\Upsilon_p} W\left\{e^{j(\theta+f_p)}\right\} \qquad (4\text{-}22)$$

where $\theta = 2\pi k/K$ (normalized frequency)

Equation (4-22) indicates that each of the interference harmonics in the spectrum is represented by a single pair of window transforms (at positive and negative frequencies f_p). With carefully chosen window shape and length (and/or sufficiently high pitch frequency), each interference harmonic can be individually resolved and the amplitudes D_p estimated by sampling the magnitude DFT at the frequencies f_p. A 40 msec Hamming window is selected, as discussed in section 4.1.3, as a a good compromise between frequency and time resolution.

The minimum size FFT required for 40 msec of data at a sampling frequency of 10 kHz is 512 points, which yields

spectral samples spaced 20 Hz apart. Unfortunately, the interference harmonics do not always occur every 20 Hz, so interpolation of the spectral values is required to obtain the most accurate amplitude estimates at the exact harmonic frequencies. A simple way of accomplishing this is by appending zeroes to the 40 msec windowed data and using a higher-order FFT (the zero-padding is strictly for interpolation purposes since the basic resolution of the spectral analysis is fixed by the 40 msec Hamming window).

Because the interference harmonic amplitudes D_p are estimated from the co-channel signal, there will be estimation errors due to the presence of the desired speech. One possible solution is to first derive the spectral parameters of the desired speech and use these to improve the estimates of D_p; however for the low SNR cases of interest here, it is very difficult to derive <u>any</u> parameters of the desired speech. Therefore, without desired speech spectral information, the best estimates of the D_p's come from the points where the interference has the highest spectral amplitudes, which are at the pitch harmonics f_p.

The noise magnitude spectrum estimate (used for spectral subtraction) is based on the estimated harmonic amplitude coefficients and the known frequency response characteristics of the Hamming window. As mentioned earlier, the length of the Hamming window has been chosen such that the

mainlobes of the pitch harmonics of the windowed noise do not usually overlap. Further, the sidelobes of the Hamming window are more than 40 dB down from the mainlobe peak and drop off at an asymptotic rate of 20 dB per decade. With this degree of selectivity, it can be assumed that the interaction between the windowed noise harmonics is minimal. Thus, given a set of estimated noise harmonic amplitudes $\hat{D}_p$, the noise magnitude spectrum can be expressed approximately in terms of only the window's mainlobe characteristic, W_{ml}, at each pitch harmonic. Replacing each window spectrum with W_{ml} in equation (4-22) (only positive values of the normalized frequency θ indicated for simplicity) then gives:

$$|\hat{N}_w(k)| = \frac{1}{2} \sum_{p=1}^{L} \hat{D}_p \, W_{ml}[\theta - f_p] \qquad (4\text{-}23)$$

where

$$W_{ml}(\theta) = \begin{cases} W(e^{j\theta}) & \text{for } |\theta| < \text{first zero of } W(e^{j\theta}) \\ 0 & \text{for } |\theta| \geq \text{first zero of } W(e^{j\theta}) \end{cases}$$

Fig. 4-6 illustrates the principle for the p-th harmonic of the noise. The interpolated "s+n" magnitude spectrum (the solid line) is evaluated at the frequency pF_o, yielding the value of $\hat{D}_p$. Then the noise magnitude is approximated by the mainlobe of the Hamming window frequency response scaled to equal $\hat{D}_p$ at its peak. This is represented by the dashed line in Fig. 4-6 (the first sidelobes are shown for

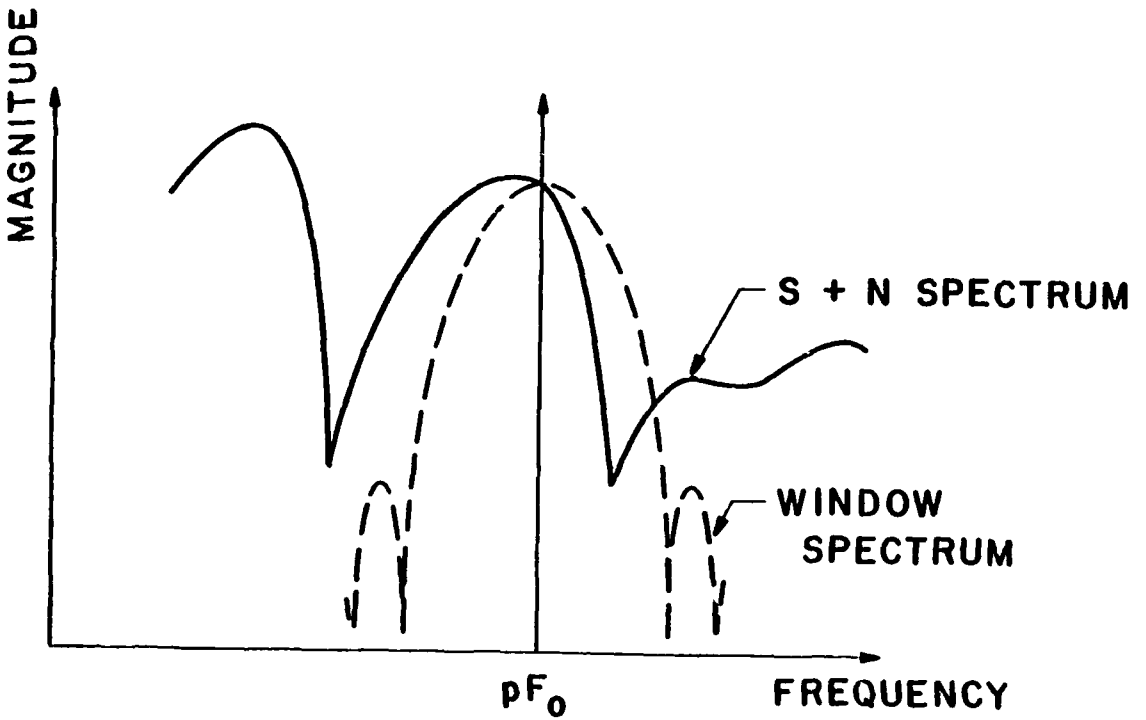


Fig. 4-6: Reconstruction of Magnitude Spectrum of p-th Interference Harmonic (from Hamming window mainlobe)

reference only and are not used in the approximation). The harmonic sampling and noise magnitude spectrum reconstruction described above provide the $|\hat{N}|$ input to the spectral magnitude subtraction, as indicated in the HMS algorithm block diagram of Fig. 4-7.

Adaptive Pitch Correlation

An adaptive pitch optimization algorithm is indicated by the dashed "feedback" from the spectral differencer to the noise pitch estimation. The purpose of this algorithm is to correct for small errors in the initial pitch estimate by perturbing the pitch until the power of the spectral difference is minimum. When the interference is of much larger amplitude than the desired speech (generally true for negative decibel SNR) and the interference signal is periodic, the power at the output of the spectral differencer should be minimized when the "true" noise pitch is attained.

Assuming most of the errors in the initial pitch estimates are only a few Hertz, the pitch perturbation procedure described above finds the pitch value which provides the most noise suppression. It should be noted that pitch errors outside the perturbation range will not be corrected. However, the perturbation range must be kept small because if it is too large, the power minimization can be affected by desired speech harmonics and/or multiples of the wrong

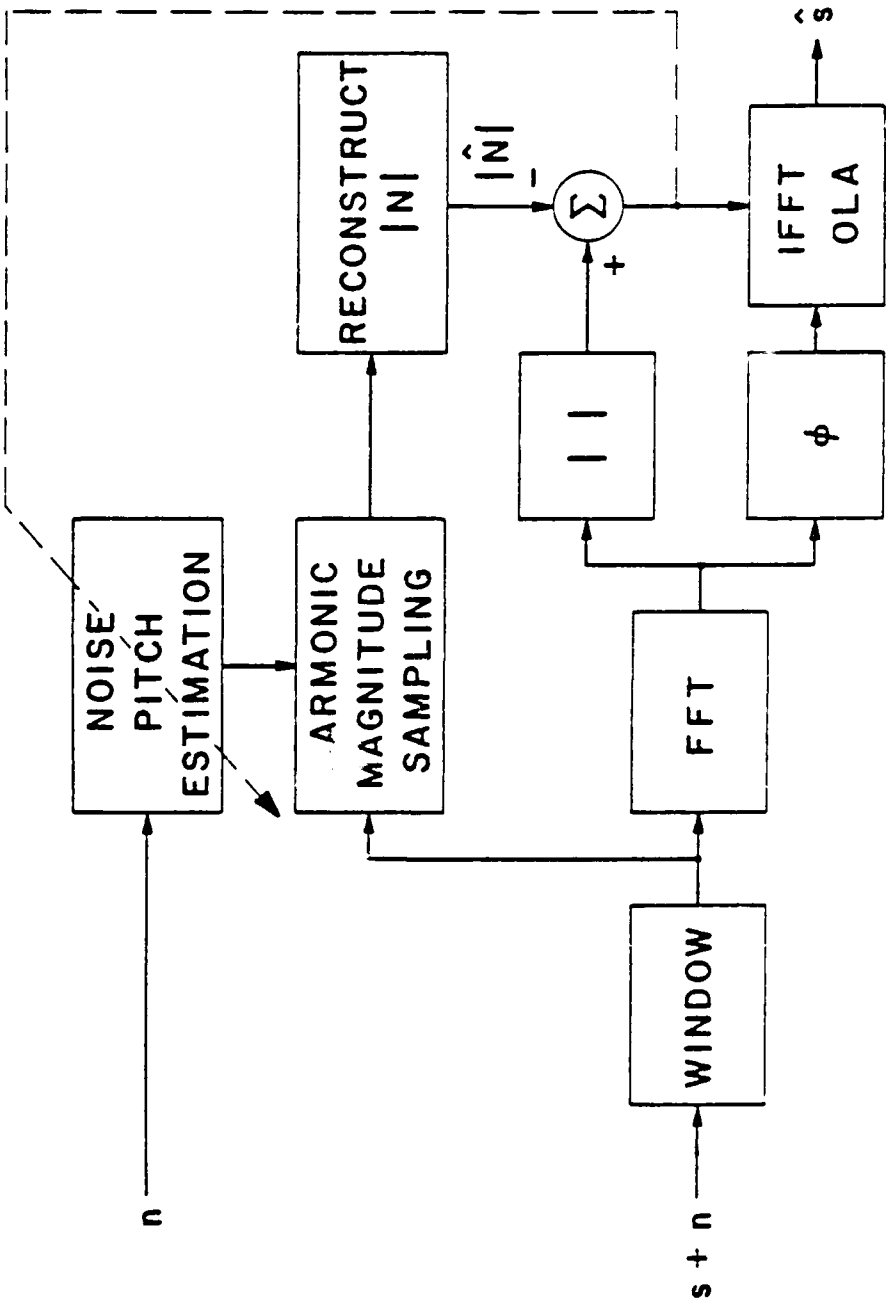


Fig. 4-7: Harmonic Magnitude Suppression (HMS) Algorithm

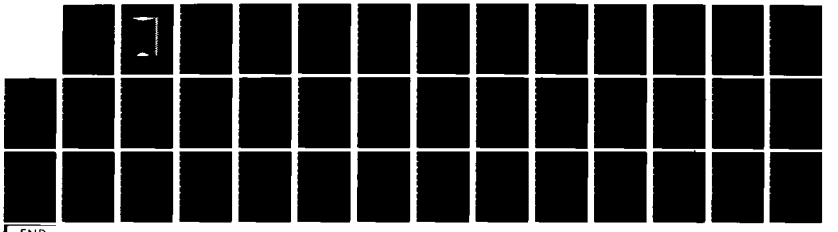
noise harmonic (i.e. the 3rd, 6th, 9th, ... pitch harmonics can be interpreted as due to $1.5F_o$ instead of F_o). A pitch perturbation of ±3 Hz (in 1 Hz increments) was found sufficient to correct most of the small errors; larger pitch perturbations did not significantly improve the SDM performance of the noise suppression algorithm.

Operation of the pitch correction routine is illustrated with an example in Fig. 4-8. The test signal consists of two harmonically-related sinusoids (at 96 and 576 Hz). The top waveform is the output from the HMS algorithm as the initial pitch estimate is swept in steps of about 1 Hz through a range of frequencies which includes the test signal fundamental of 96 Hz. When the initial pitch estimate is within ±3 Hz of the fundamental, indicated by the region between the dashed vertical lines, the test signal is almost totally suppressed. Note the difference in scales between the output and input is about nineteen to one.

The test just described illustrates the upper bounds on system performance because the test signal is perfectly periodic and there is no competing signal (i.e. no desired speech) to introduce errors into the spectral estimation. When tested on co-channel speech, the HMS algorithm provides a lesser, but still significant, amount of interference suppression. The amount of suppression depends on the accuracy of the harmonic model.

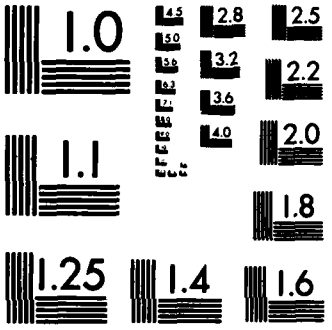
AD-A135 702 PROCESSING TECHNIQUES FOR INTELLIGIBILITY IMPROVEMENT TO SPEECH WITH CO-C.. (U) SIGNAL TECHNOLOGY INC GOLETA CA B A HANSON ET AL. SEP 83 RADC-TR-83-225 2/2

UNCLASSIFIED F30602-81-C-0226 F/G 17/2 NL



END

FILMED
DTIC



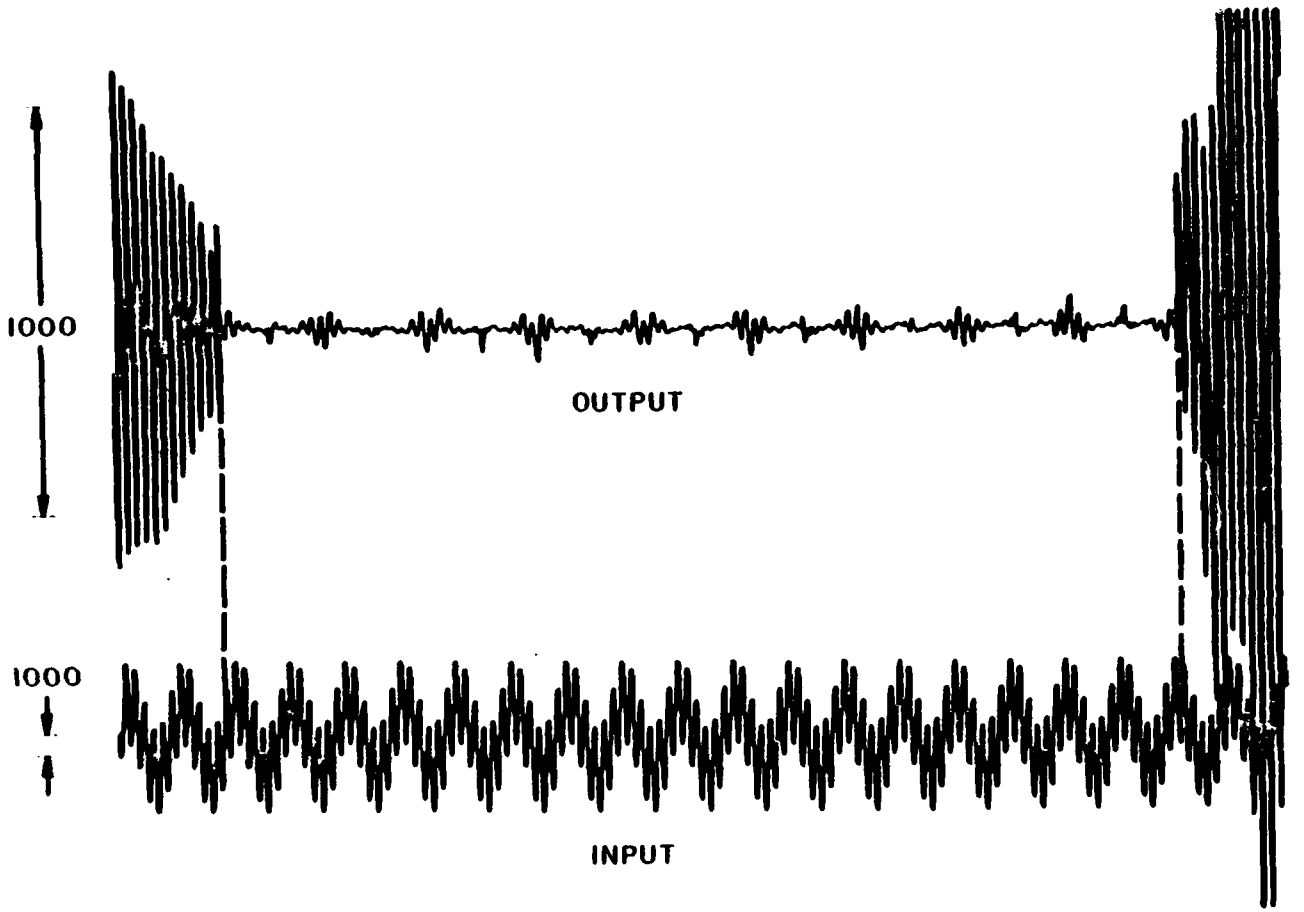


Fig. 4-8: Harmonic Magnitude Suppression (sum of sinusoids test)

Magnitude Subtraction

Power and root magnitude spectral subtraction were compared with magnitude spectral differences, and the results are summarized in Table 4-3. Similar to the tests of the SS/HS technique of section 4.2.1, all three subtraction methods gave rather close SDM's. Results for the magnitude and root magnitude cases are particularly close. Informal listening comparisons between them are consistent with the distortion performance results. The magnitude and root magnitude samples contain less perceivable interference than the power subtraction cases. However, the root magnitude method is perceived to distort the quality and gain characteristics of the desired speech more than the magnitude difference method. Thus, magnitude spectral subtraction is found to be the best approach for harmonic suppression. The HMS algorithm will be compared with the other two algorithms in the next section.

4.4 Algorithm Performance Comparisons

Three methods of noise estimation and suppression have been developed in this chapter: noise estimation using spectral sampling/harmonic synthesis (SS/HS), LPC noise synthesis (LPCN), and harmonic magnitude suppression (HMS). These algorithms are compared based on SDM calculations and informal listening evaluation. The implementations of the three algorithms were covered in the preceding two sections.

CASE	PROCESSED SAMPLE NUMBER						COMMENTS
	SDM in dB (18 dB limit SDM)						
	#3	#4	#5	#6	#7	#8	
Mag. sub [a=1]	9.1 (7.7)	9.7 (7.4)	10.1 (7.8)	10.3 (8.0)	11.7 (8.0)	8.9 (7.3)	Interference significantly reduced; a little distortion of desired speech
Power sub [a=2]	10.1 (8.1)	10.8 (7.9)	11.9 (8.5)	11.9 (8.7)	12.5 (8.4)	10.3 (8.1)	More interference than magnitude cases
$\sqrt{\text{mag sub}}$ [a=1/2]	8.8 (7.9)	9.2 (7.4)	9.3 (7.7)	9.4 (7.7)	11.3 (8.4)	8.2 (7.1)	Significant distortion of desired voice quality and gain contour

Test setup:
- -12 dB SNR
- interfering speaker is dj in samples 3, 4, and 5
- interfering speaker is jt in samples 6, 7, and 8

Table 4-3 Exponent Parameter Tests on HMS Algorithm

The important parameters of the algorithms will be briefly reviewed below.

All three algorithms are tested with "clean" pitch derived from the known interference signal. The spectral magnitude subtraction component is the same for all three approaches: the gain factor g_s is set to one, 40 msec Hamming windows are applied to the "s+n" signal before FFT, and a 20 msec window overlap is used. The SS/HS and HMS algorithms also utilize the FFT output for spectral sampling.

The LPCN algorithm applies a 200-point window with a 12th-order LPC autocorrelation analysis to the co-channel signal for estimation of the interference spectral envelope parameters. The interference synthesis is performed with pitch synchronous interpolation of the gain, pitch, and reflection coefficients [Markel and Gray 1976].

In the HMS algorithm, the pitch perturbation range is set at ±3 Hz (in 1 Hz steps). As will be shown, this small amount of pitch perturbation improves the results from the algorithm, even though the pitch contours were extracted from the "clean" interference signal. Results from the HMS algorithm without pitch perturbation (i.e. perturbation = 0 Hz) are included for comparison.

The SDM comparisons for these tests are shown in Table 4-4. As these figures indicate, the HMS algorithm with ±3 Hz pitch perturbations produces the lowest overall SDM

ALGORITHM	PROCESSED SAMPLE NUMBER SDM in dB (18 dB limit SDM)						COMMENTS
	#3	#4	#5	#6	#7	#8	
LPCN	9.6 (7.8)	9.8 (7.6)	11.5 (8.6)	11.1 (8.4)	11.9 (8.3)	9.7 (7.7)	Suppressed but "recognizable" interfering speaker
SS/HS	9.5 (8.0)	9.7 (7.5)	11.0 (8.4)	10.3 (8.0)	11.7 (8.2)	9.2 (7.5)	Interference more "noise-like" than LPCN case
HMS (Δ_{pitch}=0Hz)	9.9 (8.0)	10.3 (7.7)	11.2 (8.3)	10.9 (8.2)	12.3 (8.3)	9.8 (7.9)	Very similar to SS/HS case
HMS (Δ_{pitch}=±3Hz)	9.1 (7.7)	9.7 (7.4)	10.1 (7.8)	10.3 (8.0)	11.7 (8.0)	8.9 (7.3)	Sounds like above two cases, except interference slightly more suppressed and "noise-like"

Test setup:
- -12 dB SNR
- interfering speaker is dj in samples 3, 4, and 5
- interfering speaker is jt in samples 6, 7, and 8

Table 4-4 Spectral Distortion Measure and Informal Listening Comparisons

values for both standard and 18 dB-limited spectral distortions. However, the SDM values for all the algorithms are relatively close.

Informal listening comparisons find comparable amounts of interference suppression for all three algorithms, although there were noticeable differences in the quality of the processed outputs. The most obvious quality differences occur between the LPCN and the other processing methods. While the voiced interference remaining in the processed output of the SS/HS and HMS algorithms is considerably distorted and sounds "whispered or buzzy", the residual interference using LPCN sounds speech-like.

The difference noted in the quality of the LPCN data is also evident in the time waveform and spectral plots of the output. Comparisons of sample outputs from the LPCN and HMS (with no pitch perturbation) algorithms for a segment of co-channel speech where the desired speaker is virtually silent are shown in Figs. 4-9 and 4-10. In this case the appropriate output would be zero. While the HMS algorithm removes most of the pitch harmonics of the noise, the LPCN misses several important harmonics, so the residual interference waveform appears periodic and sounds like voiced speech. Such incomplete cancellation of the interference in the LPCN case is expected because in the LPCN algorithm the interference estimate is based on the LPC

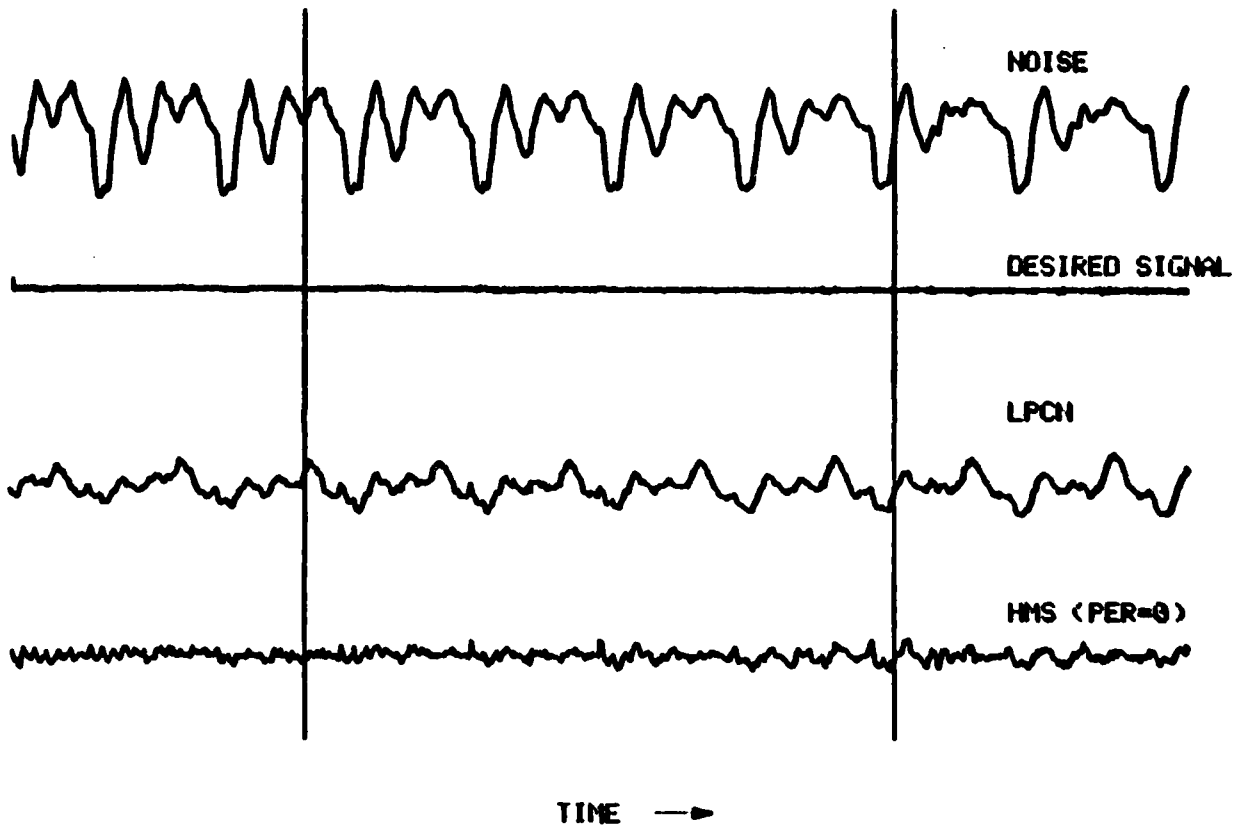


Fig. 4-9: Noise, Desired Signal, LPCN-Processed, and HMS-Processed Waveforms (segments between vertical lines used for spectra in Fig. 4-10)

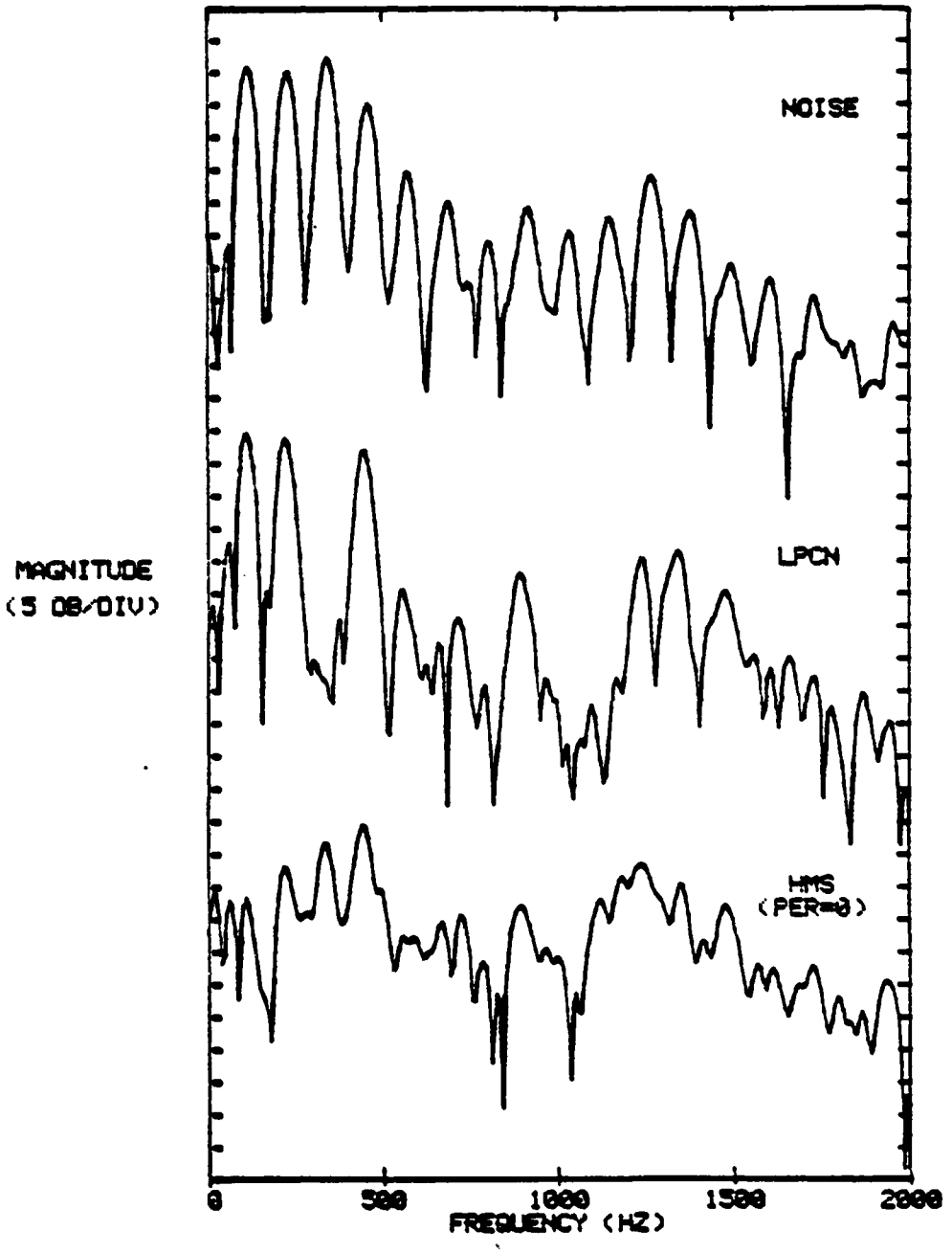


Fig. 4-10: Spectra of Noise, LPCN-Processed, and HMS-Processed Data (see Fig. 4-9)

model spectrum, which is an approximate fit to the "s+n" spectrum, while the HMS and SS/HS methods directly sample the "s+n" spectrum. This illustrates the importance of accurate interference spectrum estimation for spectral subtraction.

The HMS and SS/HS are preferred over the LPCN algorithm because the interference is not speech-like, allowing the listener to focus on the desired speaker's voice. The differences between the HMS and SS/HS algorithms are much more subtle, which is expected since both algorithms estimate the interference by spectral sampling. Without adaptive pitch correction, the HMS and SS/HS output sound very similar. The extra interference suppression obtained with adaptive pitch correction (with a ± 3 Hz perturbation range) is a small, but perceivable, improvement.

Both SDM comparison and informal listening finds the HMS with adaptive pitch correction to be the preferred approach. Formal intelligibility evaluation of the method will be discussed in chapter five.

5.0 FINAL ALGORITHM TEST AND EVALUATION

Based on the spectral distortion measure and informal listening comparisons discussed in section 4.4, the harmonic magnitude suppression (HMS) algorithm was selected for the final intelligibility test. The HMS algorithm tested is briefly summarized in section 5.1. The test procedures are discussed in section 5.2. The results are presented in section 5.3.

5.1 The Harmonic Magnitude Suppression (HMS) Algorithm

A block diagram of the processing algorithm tested is shown in Fig. 5-1. It is the HMS algorithm, discussed in sections 4.3 and 4.4 except for one small change. The change is to use maximum estimated noise power instead of minimum spectral difference power as the feedback for pitch correction. This is shown in Fig. 5-1, where the dashed line (indicating the feedback path for pitch correction) originates from the estimated noise spectrum, $|\hat{N}|$, instead of from the spectral magnitude difference $|\hat{S}|$, as shown in Fig. 4-7. The new feedback produced equivalent SDM results and the speech quality is informally judged to be the same. This change saves computation time by avoiding the square-root (required for magnitude subtraction) until all the pitch perturbations are finished.

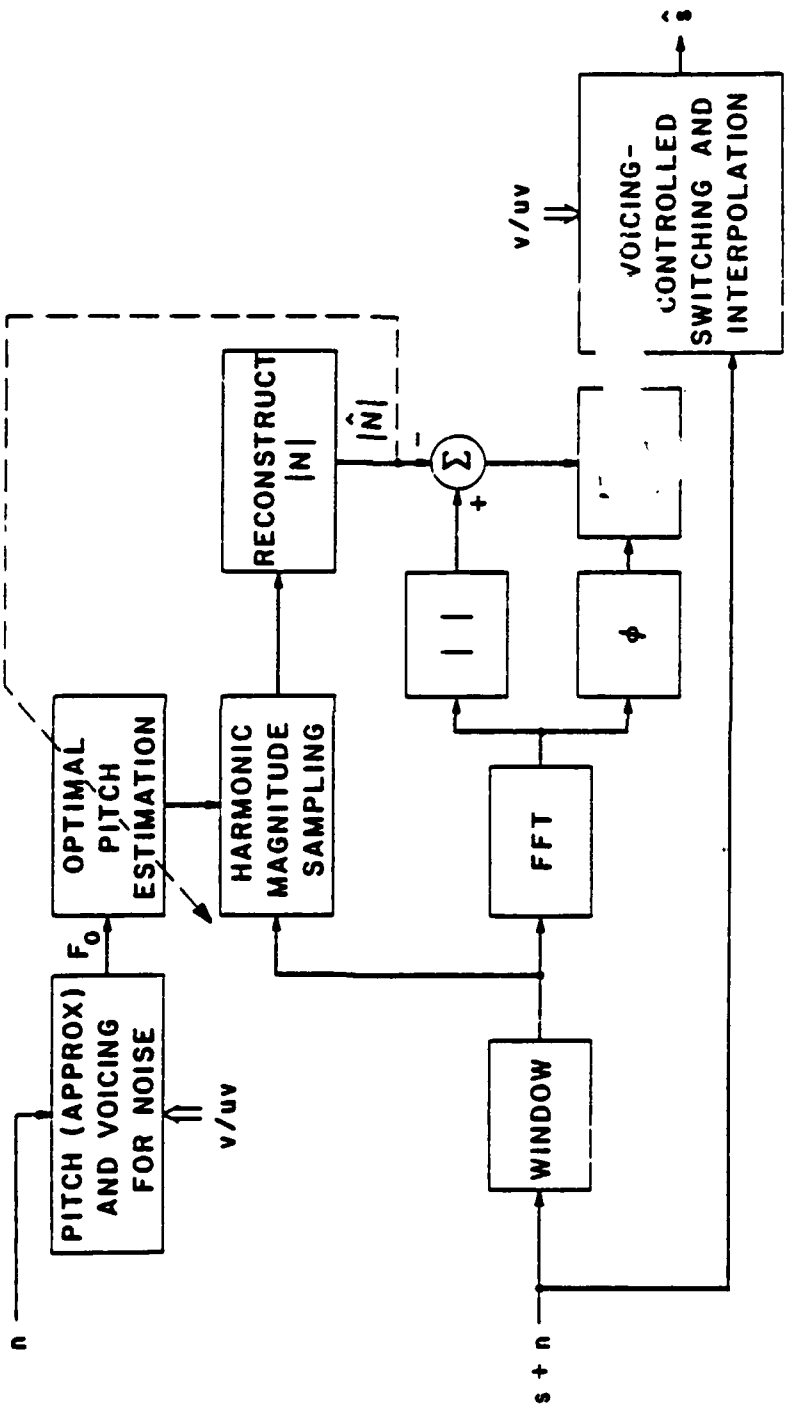


Fig. 5-1: Final Test System

The pitch and voicing parameters used to estimate the interference are extracted from the known interference, as indicated by the noise input into the pitch and voicing box. The assumption of "clean" pitch and voicing information has been used throughout this work (and in previous studies [Perlmutter et al. 1977]) for testing. This allows separation of the pitch detection problem from the HMS algorithm. Except for this assumption on pitch and voicing, the rest of the system of Fig. 5-1 is realizable and requires no other a priori information about the co-channel signal. It should be emphasized again that for co-channel speech with low SNR's (i.e. -6 and -12 dB) tested in this study, accurate pitch and voicing estimation for the interference signal is reasonably achievable because the interference is generally much stronger than the desired signal.

The HMS algorithm applies only to voiced interference segments. The unvoiced segments are passed through. It should be noted that this approach occasionally leads to distractingly high levels of unvoiced interference in the output. One-frame linear interpolation between processed and unprocessed data is performed at voicing transitions to avoid abrupt changes. This is shown as the "voicing-controlled switching and interpolation" in Fig. 5-1.

5.2 Intelligibility Testing

Details of the testing procedures have been discussed in section 2.1. Only several points specific to this test are discussed here. They are summarized in Table 5-1.

The first three items in Table 5-1 relate to test data preparation. Ten phonetically balanced sentences were used for the desired speaker, and ten different PB sentences for the interference (split evenly between two different interfering speakers). The text of the test sentences is included in appendix B. Co-channel test data with SNR's of -6 and -12 dB was constructed from these sentences using the procedures described in section 2.1.

The listener panel consists of ten subjects, seven of which were professionals or graduate students in the speech and hearing field. Trained listeners were selected on the assumption that they would yield more consistent results, which is generally verified by the results. All the listeners had no prior experience with co-channel type data, and thus required some orientation and training, as discussed in section 2.1, by way of a handout (appendix A) and short demonstrations.

The HMS-processed data was tested as an enhancement to the unprocessed co-channel data. That is, the subjects heard processed and unprocessed data for half of the sentences, and only unprocessed data for the other half of the

- 1 Desired Speaker, Two Different Interfering Speakers
- Ten Phonetically Balanced Desired Speaker Sentences and Ten Phonetically Balanced Interfering Speaker Sentences
- -6 dB and -12 dB SNR's
- Ten Listening Subjects
- Unprocessed Only: 5 Sentences
  Unprocessed and Processed: 5 Sentences
- Multiple Listens Allowed
- Orthographic Transcription

Table 5-1 Final Intelligibility Tests

sentences (this is the intelligibility improvement test procedure discussed in section 2.1). All listening subjects heard the -12 dB data first. For each speech sample, as many repeats as needed were allowed. After a short break, the subjects were presented the -6 dB test. The data was presented in the same order as the earlier test. It was assumed that since the data at -6 dB would be more intelligible than in the -12 dB test, and as many listens as needed were allowed, the later session (-6 dB) did not benefit from the earlier one (-12 dB). At the end of both listening sessions, the transcriptions were scored. The results are discussed in the next section.

5.3 Results and Analysis

The listener transcriptions are scored according to the rules defined in section 2.1. The results are tabulated in Table 5-2. Each entry in the table is the number of words a subject correctly (or partially) transcribed from a sample. The even numbered subjects in the table heard the even numbered sentences after processing, and the odd numbered subjects heard the the odd numbered sentences after processing. Thus each sentence was heard by five subjects after processing, and by the other five without processing.

The average intelligibility scores are computed to provide an overall evaluation of the enhancement algorithm. First the probabilities of correctly transcribing a word

SUBJECT (Init;ID#)		SPEECH SAMPLE NUMBER (Scores)										
		#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	
BW	2	0.	4.00	0.	2.00	3.50	6.00	7.00	7.00	8.00	7.00	
TA	3	0.	0.50	0.	1.00	5.00	4.00	4.00	4.25	7.33	2.25	
DB	4	0.	0.	1.00	0.	4.00	6.00	2.00	7.00	6.00	1.50	
MB	5	2.00	1.00	2.00	0.	4.00	4.00	4.00	4.00	1.50	6.50	
PF	6	0.	5.50	1.50	3.00	2.50	6.00	3.00	7.00	6.00	1.00	-12 dB SNR test
TJ	7	0.	5.00	6.00	6.00	5.00	4.00	5.00	7.00	8.00	4.00	
NP	8	1.00	6.00	5.00	4.00	5.00	4.00	2.00	4.50	7.00	4.00	
LA	9	3.00	4.00	3.00	6.00	5.00	5.00	1.50	7.00	8.00	6.50	
DS	10	0.	2.00	1.00	6.00	1.50	3.00	4.00	7.00	3.00	5.00	
BD	11	0.	3.00	6.00	2.00	5.00	2.50	6.00	7.00	7.00	7.00	
BW	2	7.00	6.00	6.00	4.50	3.50	6.00	7.00	7.00	8.00	7.00	
TA	3	4.00	6.00	5.00	6.00	6.00	4.00	5.50	6.50	8.00	6.00	
DB	4	0.	2.50	5.00	0.	4.00	6.00	5.00	7.00	6.00	3.50	
MB	5	2.00	3.00	5.00	2.00	4.00	4.00	5.00	7.00	8.00	6.50	
PF	6	2.00	6.00	6.00	4.50	5.00	6.00	7.00	7.00	8.00	6.00	-6 dB SNR test
TJ	7	2.00	6.00	6.00	6.00	5.00	4.00	6.00	7.00	8.00	7.00	
NP	8	2.00	6.00	5.00	4.00	4.00	6.00	6.00	7.00	8.00	7.00	
LA	9	3.00	6.00	6.00	5.50	6.00	5.00	5.00	7.00	8.00	6.00	
DS	10	0.	5.50	1.00	6.00	3.50	3.00	5.00	7.00	5.00	5.50	
BD	11	2.00	6.00	6.00	6.00	5.00	3.00	6.00	7.00	8.00	7.00	

(Even numbered subjects heard even sentences with processing and odd numbered subjects heard odd sentences with processing)

Table 5-2: Intelligibility Test Scores

with and without processing are estimated:

$$\hat{p}_p = \frac{\text{number of words correct with processing}}{N_p} \qquad (5\text{-}1)$$

$$\hat{p}_u = \frac{\text{number of words correct without processing}}{N_u} \qquad (5\text{-}2)$$

where

N_p = total number of words presented with processing

N_u = total number of words presented without processing

The average intelligibility improvement is then defined as the difference, $\Delta = \hat{p}_p - \hat{p}_u$, of the above. The calculated values, expressed in percentages for the SNR's of -6 and -12 dB, are given in Table 5-3. The most important result shown there is an increase in intelligibility for the -12 dB case from 53.8% without processing to 62.7% with processing. This 8.9% intelligibility increase means 17% more words became intelligible after processing. The improvement of 3.6% for the -6 dB test is considerably smaller, but this was expected since the initial intelligibility for unprocessed speech is 78.3%, leaving little room for improvement.

Confidence levels of the intelligibility gain were computed based on the following statistical model of the test. It is first assumed that the test procedure has removed as much of the biases and variation as possible from the exper-

CASE	SNR (dB)	SUBJECTS (number)	WITHOUT PROCESSING ($\hat{p}_u$ x 100%)	WITH PROCESSING ($\hat{p}_p$ x 100%)	PERCENTAGE IMPROVEMENT (Δx100%)
A	-12	all (10)	53.8	62.7	8.9
B	-12	"top scorers" (6)	64.4	70.5	6.1
C	-6	all (10)	78.3	82.0	3.6
D	-6	"top scorers" (6)	87.1	87.6	0.5

Table 5-3 Intelligibility Scores (% Correct)

iment so that only the variable of interest (intelligibility) affects the final results. Secondly, assume that a word transcribed from the unprocessed data can be either correct (with a probability of p_u) or wrong (with a probability of $q_u = 1 - p_u$). Then if the probability p_u is assumed to be the same for all of the unprocessed words, a transcription of each word can be considered a Bernoulli trial. A shortcoming of the model is that the probability of a listener correctly transcribing each word is independent of all the other words transcribed in the test (by himself or other listeners). With the above assumptions the total number of correct transcriptions for a particular data condition (processed or unprocessed) has a binomial distribution. The transcriptions for data with and without processing are thus two different binomial processes. The mean and standard deviation of the difference between the probabilities of correct transcription can be estimated by:

$$\mu_\Delta = \hat{p}_p - \hat{p}_\mu \tag{5-3}$$

$$\sigma_\Delta = \sqrt{\frac{\hat{p}_p\hat{q}_p}{N_p} + \frac{\hat{p}_u\hat{q}_u}{N_u}} \tag{5-4}$$

Given this statistical model, it is possible to test the "null hypothesis": that $p_p = p_u = p$. For sufficiently large N's (i.e. $Npq \geq 9$ [Siegel 1956]), the probability

differences approach a Gaussian distribution. Substituting p for $\hat{p}_p$ and $\hat{p}_u$ into equations (5-3) and (5-4), this Gaussian distribution can be expressed in terms of the standardized variable z:

$$z = \frac{\hat{p}_p - \hat{p}_u}{\sigma_\Delta} \qquad (5\text{-}5)$$

where

$$p = \frac{N_p\hat{p}_p + N_u\hat{p}_u}{N_p + N_u}$$ [estimated probability under null hypothesis: $p_p = p_u = p$]

With the above formulation, the level of confidence that the null hypothesis is false (i.e. the difference between $\hat{p}_p$ and $\hat{p}_u$ is not due to chance) can be calculated. A "one-tailed" test of the hypothesis assumes in this case that processing only adds information, and gives the confidence level for $p_p > p_u$ (i.e. including the processed data gives a higher probability of a correct transcription than using unprocessed data alone). The critical value, z_c, for the above distribution is obtained by substituting the estimated probabilities $\hat{p}_p$ and $\hat{p}_u$ into equation (5-5). Then the Gaussian variable z is integrated from z_c to infinity, providing the probability of rejecting the null hypothesis. The level of confidence, L_{conf}, in the hypothesis that $P_p > p_u$, is defined as:

$$L_{conf} = 1.0 - \frac{1}{\sqrt{2\pi}} \int_{z_c}^{\infty} e^{-z^2/2}\, dz \qquad (5\text{-}6)$$

Tabulated values of the above integral versus z_c are readily available [e.g. Siegel 1956 and Spiegel 1961]. From the correct transcription percentages given for the -12 dB case above, the confidence level for the hypothesis that processed speech improves the intelligibility is over 98%.

The intelligibility scores at -12 and -6 dB SNR (cases A and C of Table 5-3) are plotted in Fig. 5-2. The solid line for unprocessed speech is provided for reference; the distance between this line and points A and C gives the intelligibility gain with processing. Although only two SNR values were used in these tests, approximate intelligibility gains at other points can be estimated by retabulating the data. For example, if only the top scoring listeners in each test are considered, the intelligibilities of cases B and D in Table 5-3 and Fig. 5-2 are obtained. These were calculated by separately ranking the even and odd numbered subjects and selecting the top three of each group as "top scorers". Six subjects were chosen for these groups because of the separation of their scores from the lowest three or four scores.

To extrapolate points A through D to lower intelligibility values, another approach is taken. The data from the -12 dB SNR test are ranked sentence-by-sentence according to their intelligibility without processing. Then the most intelligible sentences are successively removed, and the

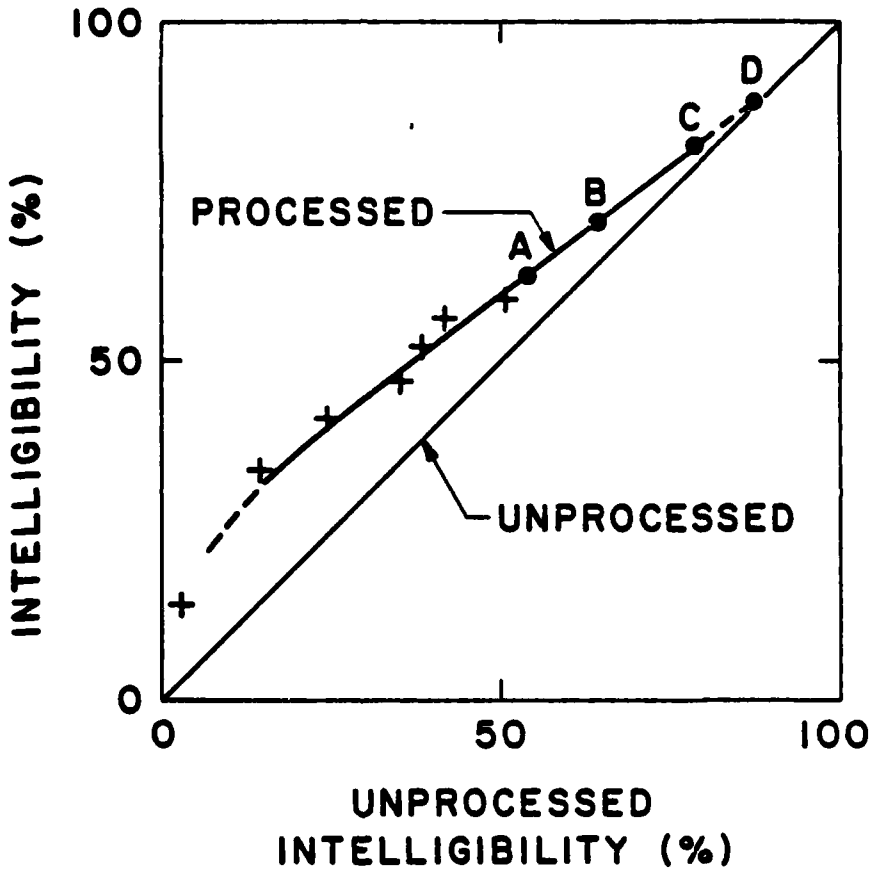


Fig. 5-2: Processed and Unprocessed Intelligibility

intelligibilities with and without processing are recalculated for the remaining data; sentences with intelligibilities within a few percentage points are removed together, otherwise the removal of data is one sentence at a time. The intelligibility values obtained in this manner are indicated by the crosses in Fig. 5-2. It should be noted that the amount of data used to calculate each point decreases with the intelligibility (the point nearest the origin represents only one sentence); the confidence assigned to these points decreases accordingly.

The increase in intelligibility gain with decreasing unprocessed intelligibility shown in Fig. 5-2 is even more apparent in Fig. 5-3, which plots relative intelligibility improvement (gain ÷ unprocessed intelligibility). The one standard deviation limits for each point (based on the Gaussian approximations) illustrate the increase in score variability as fewer sentences are included.

The trend indicated by Figs. 5-2 and 5-3 is very significant: the gain from the HMS processing appears to increase, up to a limit, as the intelligibility for unprocessed data (and by implication SNR) decreases. Such behavior can be explained as follows. The accuracy of the estimated noise parameters (pitch and harmonic amplitudes for the HMS algorithm) increases as SNR decreases, the noise suppression improves, and thus the intelligibility gain increases.

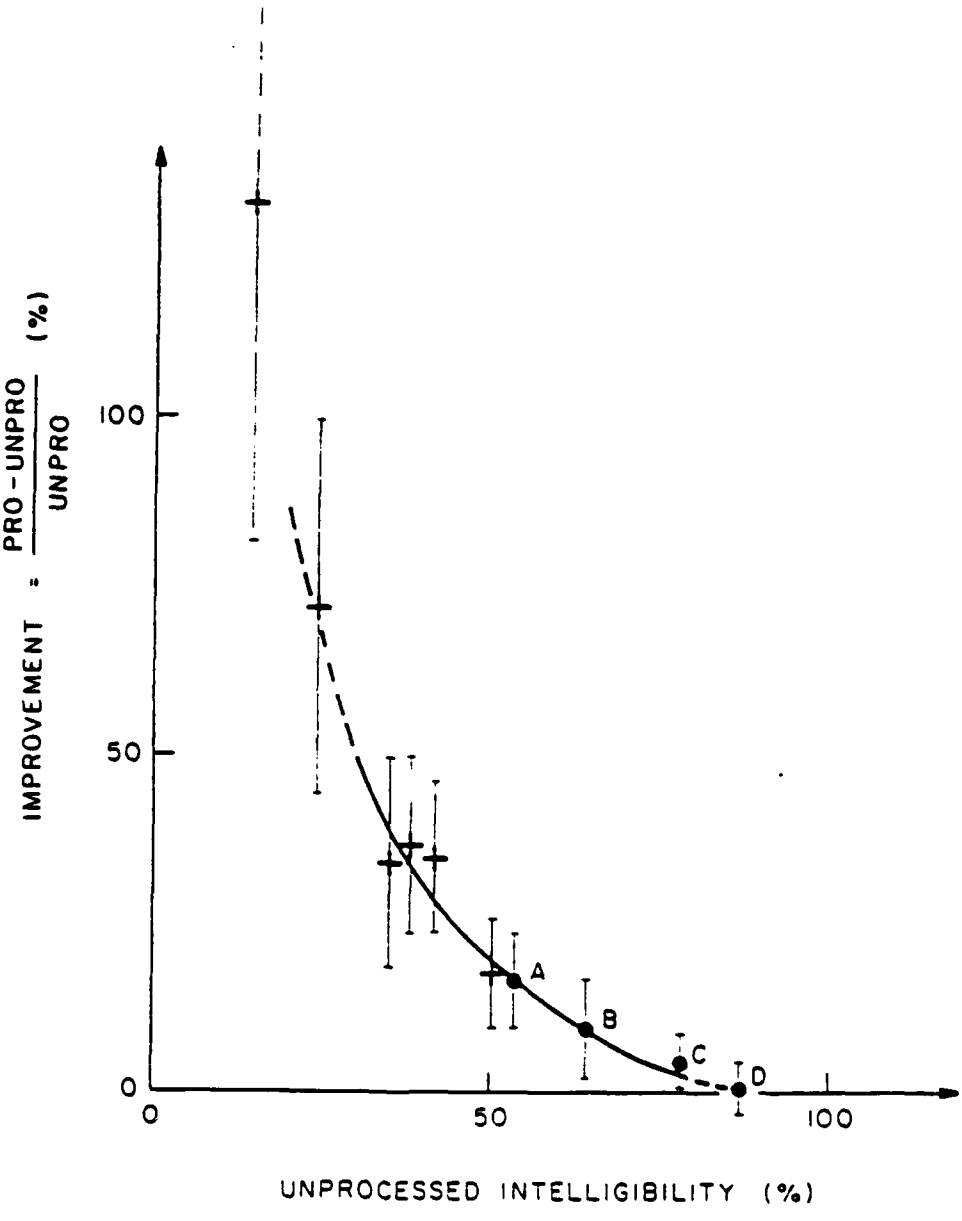


Fig. 5-3: Relative Intelligibility Improvement (with one standard deviation intervals)

6.0 CONCLUSIONS AND RECOMMENDATIONS

6.1 Conclusions

Several post-processing techniques for separating co-channel speech have been studied and tested in this research. The major conclusions derived from this work are:

1) The harmonic magnitude suppression (HMS) technique significantly improves intelligibility for SNR < -6 dB.

This is the key result of the research. As reported in chapter five, for -12 dB SNR co-channel data, an increase in intelligibility from 53.8% before processing to 62.7% for the cases with processing (representing a percentage gain of 17% more words) was obtained. Statistical analysis of the test data shows the result to be valid at a 98% confidence level. No previous research in this area has demonstrated any measurable intelligibility gains.

2) Intelligibility improvement with HMS processing generally increases as SNR decreases.

Further analysis of the intelligibility test data, as discussed in section 5.3, has shown that the relative intelligibility gain tends to increase as the unprocessed intelligibility (and SNR) decreases. In other words, the HMS technique is most effective for the most corrupted data.

3) The signal extraction algorithm based on harmonic synthesis does not improve intelligibility.

While the test results on data processed with the harmonic synthesis extraction approach of chapter three indicate that no intelligibility improvement was obtained, this initial work provided several new directions for investigation.

4) The potential for intelligibility improvement is highest for signals with SNR < 0 dB. This leads to a SNR-dependent processing concept.

The emphasis on negative decibel SNR cases, derived from the initial intelligibility tests, concentrated the effort on interference suppression (the logical approach for SNR < 0 dB), which ultimately led to the successful HMS technique. Further, the importance of the zero decibel threshold for "instantaneous" SNR suggests that SNR control of the processing is a promising concept that deserves close study.

5) The spectral distortion measures (SDM) developed are found to be useful algorithm development tools.

Algorithm performance measurement with SDM's provides a useful alternative to the often unreliable evaluations of informal listening, and helps reduce dependence on time-consuming formal intelligibility testing. It should be

emphasized that SDM evaluation of processed co-channel speech is a new concept, and until formal studies determine a more exact relationship to co-channel speech intelligibility, SDM evaluation should only be used as a developmental tool and not as a replacement for final formal intelligibility testing of algorithm performance.

6.2 Recommendations

The research which resulted in the intelligibility gains reported here represents significant progress towards realization of a useful co-channel speech separation system. To further develop this system, the following research directions are recommended:

1) Automatic pitch and voicing detection

The implementation of automatic pitch and voicing detection is the key item remaining for completion of the suppression system. This is a reasonable research task for the negative SNR cases of interest because the interference is of much larger amplitude than the desired signal.

2) Processing of unvoiced interference

In the HMS algorithm tested here, no processing is done when the interference is unvoiced. Although unvoiced speech is generally of lower energy than voiced speech, the unvoiced segments of the interfering speech are perceived as

much louder after the voiced segments have been suppressed.

3) SNR-dependent processing

The results presented in chapter three on the harmonic synthesis extraction method suggest that a SNR-dependent algorithm may improve overall intelligibility; with this approach, the processing is applied only on those segments with the most interference, so that possible distortions to segments with good SNR are avoided.

4) Interactive playback selection

Based on our experience with the intelligibility tests conducted for this study, we have found that interactive playback selection is desirable to allow the listeners to select between the processed and unprocessed data when both are available. Such a processed data/unprocessed data switch is recommended for use in future intelligibility tests and actual operating environments. This interactive input from the user is indicated in Fig. 6-1, which shows how the processing elements discussed above would be combined in a complete noise suppression system.

5) Performance evaluation

Intelligibility testing is recommended at each major stage of future development. This is necessary to quantify the gains obtained and to identify areas requiring more work. It is recommended that in future intelligibility

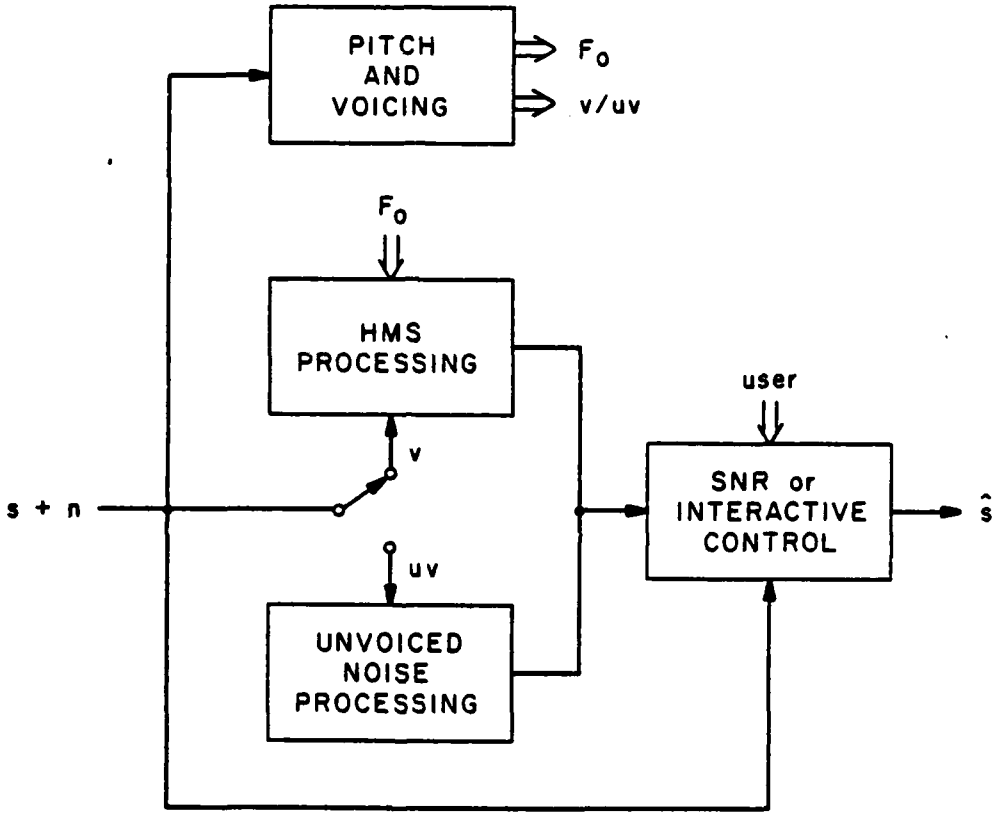


Fig. 6-1: Complete Noise Suppression System

tests more listening subjects, different talkers, and a larger range of SNR's be included. In between major steps in the development, spectral distortion measure evaluation is recommended for algorithm verification and tuning.

6) Spectral distortion measures

The utility of SDM's in evaluating co-channel separation techniques has been demonstrated in this research. A better understanding of the relationship of these measures to intelligibility is desirable in order to fully exploit their potential and further expedite algorithm development.

7) Application to automatic speech recognition

Once a prototype co-channel speech enhancement system is developed, application as a front-end to automatic speech recognition systems can be evaluated. Research efforts so far have been focused on aiding human listeners, thus the system's capabilities for improving ASR performance are yet unknown. Modifications to the co-channel separation system may be necessary to obtain optimum performance as an ASR front-end.

8) Real-time system implementation

The HMS algorithm developed is implementable in real-time with available signal processers. Thus, there are no inherent problems with developing a real-time system as long as the additional components developed for the system (i.e.

pitch-voicing detection, unvoiced interference processing, and SNR-dependent algorithm control) are also designed for real-time implementation.

APPENDIX A: Test Description Handout for Intelligibility Test Subjects

CO-CHANNEL EXPERIMENT

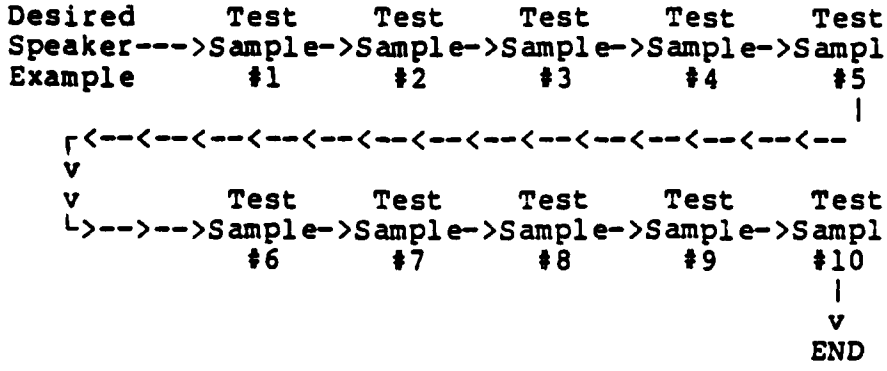
A. Purpose of Test: To evaluate relative intelligibility of different cochannel speech processing methods.

B. Test Procedure:

1) This experiment is semi-automated: proceed at your own rate. The program waits for your responses at each step of the test.

2) The test consists of ten speech samples from which it is desired to transcribe the "desired" speaker's words.

a) A "clean" example of the desired speaker will be played first; it can also be repeated at any time as a reminder of the "desired" speaker's voice.
b) If you requested no repeats or examples, then the test would have the following sequence:



c) While the "desired" speaker remains the same throughout, the interfering speaker will change after sample #5.

3) The test objective is to correctly write down what the "desired" speaker says. Note that:

a) All words are standard English.
b) Homonym spellings are acceptable (do not worry if you heard "to, too or two")
c) Plurals are important (write the plural form if you

heard it that way)
d) The articles "the, a or an" are not scored; don't worry about recording them (you can write them down if this helps)
e) Word order is important, so write down what you hear in the right order (even if it doesn't make much sense).
f) Avoid contractions (for example, do not write he's for "he is")
g) Educated "guesses" are acceptable as long as they are based on what you heard. Also, parts of words or a couple choices (such as "cup" or "sup" if you could not decide between them) can also be recorded.

4) This test is designed to be difficult, so it is easy to confuse the "desired" speaker with the interference. If you have the slightest doubt about which voice is the "desired" speaker, then record what both speakers are saying, and indicate which text is your best estimate of the desired one.

5) While there are an unlimited number of repetitions allowed, listeners generally reach a point of diminishing returns beyond which little further information can be obtained (at about 10 to 15 repetitions), so don't waste an inordinate amount of time on any single sample.

6) Also note that there is no "backtracking" feature, so previous samples cannot be reviewed! However, if you unintentionally proceed to the next sample, the missing repeats can be played at the end with help from the test co-ordinator.

7) Before starting the test several examples of processed data will be played and explained.

Sentence Number	Desired Speaker Sentence (Interfering Speaker Sentence)	Number of Scored Words
1	fairy tales should be fun to write (steam hissed from the broken valve)	7
2	we admire and love a good cook (the new girl was fired today at noon)	6
3	a young child should not suffer fright (they felt gay when the ship arrived in port)	6
4	acid burns holes in wool cloth (add the store's account to the last cent)	6
5	there the flood mark is ten inches (the sky that morning was clear and bright blue)	6
6	add the column and put the sum here (sunday is the best part of the week)	6
7	the third act was dull and tired the players (torn scraps littered the stone floor)	7
8	she has a smart way of wearing clothes (the child almost hurt the small dog)	7
9	he carved a head from the round block of marble (there was a sound of dry leaves outside)	8
10	eight miles of woodland burned to waste (the doctor cured him with these pills)	7

Desired Speaker: sw (sentences 1-10)
Interference: dj (sentences 1-5)
jt (sentences 6-10)

APPENDIX B: Final Intelligibility Test PB Sentences

REFERENCES

J.B. Allen, "Short-Term Spectral Analysis, Synthesis, and Modification by Discrete Fourier Transform", IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-25, No. 3, pp. 235-238, June 1977.

J.B. Allen, "Applications of the Short Time Fourier Transform to Speech Processing and Spectral Analysis", IEEE ICASSP-82 Proceedings, Vol. 2, pp. 1012-1015, 1982.

"Methods for the Calculation of the Articulation Index", S3.5, 1969.

C.I. Berlin & M.R. McNeil, "Dichotic Listening", in Contemporary Issues in Experimental Phonetics, (N.J. Lass, ed.), Academic Press, New York, San Francisco, London, pp. 327-387, 1976.

M. Berouti, R. Schwartz, & J. Makhoul, "Enhancement of Speech Corrupted by Acoustic Noise", IEEE ICASSP-79 Conf. Rec., pp. 208-211, Apr. 1979.

S.F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-27, No. 2, pp. 113-120, Apr. 1979.

S.F. Boll & R.E. Wohlford, "Event Driven Speech Enhancement", IEEE ICASSP-83 Proceedings, Vol. 3, pp. 1152-1155, Apr. 1983.

R.C. Cox & D.M. Robinson, "Some Notes on Phase in Speech Signals", IEEE ICASSP-80 Proceedings, Vol. 1, pp. 150-153, 1980.

R.A. Curtis & R.J. Niederjohn, "An Investigation of Several Frequency-Domain Processing Methods for Enhancing the Intelligibility of Speech in Wideband Random Noise", IEEE ICASSP-78 Conf. Rec., pp. 602-605, 1978.

S.B. Davis & P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-28, No. 4, pp. 357-366, Aug. 1980.

R.J. Dick, "Co-Channel Interference Separation", Rome Air Development Center, Report RADC-TR-80-365, Dec. 1980.

J.P. Egan, "Articulation Testing Methods", Laryngoscope, pp. 955-981, 1948.

J.K. Everton, Sr., "The Separation of the Voice Signals of Simultaneous Speakers", Ph.D. Thesis, Dept. of Computer Science, Univ. of Utah, 1975.

J.L. Flanagan & R.M. Golden, "Phase Vocoder", The Bell System Technical Journal, pp. 1493-1509, Nov. 1966.

J.L. Flanagan, Speech Analysis Synthesis and Perception, Springer-Verlag, New York, Heidelberg, Berlin, 1972.

A.J. Fourcin, W.A. Ainsworth, G.C.M. Fant, O. Fujimura, H. Fujisaki, W.J. Hess, J.N. Holmes, F. Itakura, M.R. Schroeder, & H.W. Strube, "Speech Processing by Man and Machine-Group Report", in Recognition of Complex Acoustic Signals, (T.H. Bullock, ed.), Berlin, Dahlem Workshop, pp. 307-351, 1977.

R.H. Frazier, "An Adaptive Filtering Approach Toward Speech Enhancement", S.M. Thesis, Dept. of Electrical Engineering, M.I.T., Cambridge, 1975.

S.A. Gelfand, Hearing: An Introduction to Psychological and Physiological Acoustics, Marcel Dekker, Inc., New York, pp. 239-258, 1981.

A.H. Gray, Jr. & J.D. Markel, "Distance Measures for Speech Processing", IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-24, No. 5, pp. 380-391, Oct. 1976.

R.M. Gray, A. Buzo, A.H. Gray, Jr., & Y. Matsuyama, "Distortion Measures for Speech Processing", IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-28, No. 4, pp. 367-376, Aug. 1980.

J.D. Harris, Psychoacoustics, The Bobbs-Merrill Company, Inc., Indianapolis, Indiana, pp. 24-38, 1974.

M.E. Hawley (ed.), Speech Intelligibility and Speaker Recognition, Dowden, Hutchinson & Ross, Inc., Stroudsburg, PA., 1977.

M.H.L. Hecker, G. Von Bismarck & C.E. Williams, "Automatic Evaluation of Time-Varying Communication Systems", IEEE Trans. Audio Electroacoust., Vol. AU-16, pp. 100-106, 1968.

L. Hoy, B. Burns, D. Solcan & R. Yarlagadda, "Noise Suppression Methods for Speech Applications", IEEE ICASSP-83 Proceedings, Vol. 3, pp. 1133-1136, May 1983.

A.S. House, C.E. Williams, M.H.L. Hecker, & K.D. Kryter, "Articulation-Testing Methods: Consonantal Differentiation

with a Closed-Response Set", J. Acoust. Soc. Amer., Vol. 37, No. 1, pp. 158-166, Jan. 1965.

Institute of Electrical and Electronics Engineers, "IEEE Recommended Practice for Speech Quality Measurements", IEEE No. 297, New York, June 1969.

N.S. Jayant & P. Noll, "Digital Coding of Waveforms: Principles and Applications to Speech and Video", Bell Labs (unpublished manuscript), 1982.

B.H. Juang, "Co-Channel Interference Suppression/Speech Separation", Progress Report, Dec. 1981.

K.D. Kryter, "Methods for the Calculation and Use of the Articulation Index", J. Acoust. Soc. Amer., Vol. 34, No. 11, pp. 1689-1697, Nov. 1962a.

K.D. Kryter, "Validation of the Articulation Index", J. Acoust. Soc. Amer., Vol. 34, No. 11, pp. 1698-1702, Nov. 1962b.

J.S. Lim, "Evaluation of a Correlation Subtraction Method for Enhancing Speech Degraded by Additive White Noise", IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-26, No. 5, pp. 471-472, Oct. 1978.

J.S. Lim & A.V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech", Proceedings of the IEEE, Vol. 67, No. 4, pp. 1586-1604, Dec. 1979.

J.S. Lim (ed.), Speech Enhancement, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1983.

J.D. Markel & A.H. Gray, Jr., Linear Prediction of Speech, Springer-Verlag, Berlin, Heidelberg, New York, 1976.

J.D. Markel & A.H. Gray, Jr., "Harmonic Synthesis - An Improvement to the All-Pole Synthesizer with Increasing Fundamental Frequency", STI Technical Memo, 1978.

D.W. Martin, "Uniform Speech-Peak Clipping in a Uniform Signal-To-Noise Spectrum Ratio", J. Acoust. Soc. Amer., Vol. 22, No. 5, pp. 614-621, Sept. 1950.

R.J. McAulay & M.L. Malpass, "Speech Enhancement Using a Soft-Decision Noise Suppression Filter", IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-28, No. 2, pp. 137-145, Apr. 1980.

E.E. Milios & A.V. Oppenheim, "The Phase-only Version of the

LPC Residual in Speech Coding", ICASSP-83 Proceedings, Vol. 2, pp. 797-799, 1983.

G.A. Miller, "The Masking of Speech", Psychological Bulletin, Vol. 44, No. 2, pp. 105-129, Mar. 1947.

H. Nawab, A.V. Oppenheim, & J.S. Lim, "Improved Spectral Subtraction for Signal Restoration", IEEE ICASSP-81 Proceedings, Vol. 3, pp. 1105-1108, 1981.

P. Noll, "Adaptive Quantization in Speech Coding Systems", Proc. Int. Zurich Seminar on Digital Communications, pp. B3.1 to B3.6, Oct. 1974.

T.W. Parsons, "Separation of Simultaneous Vocalic Utterances of Two Talkers", Ph.D. Dissertation, Polytechnic Institute of New York, June 1975.

T.W. Parsons & M.R. Weiss, "Enhancing/Intelligibility of Speech in Noisy or Multi-Talker Environments", Rome Air Development Center, Report RADC-TR-75-155, June 1975.

T.W. Parsons, "Separation of Speech from Interfering Speech by Means of Harmonic Selection", J. Acoust. Soc. Amer., Vol. 60, No. 4, pp. 911-918, Oct. 1976.

T.W. Parsons, "Study and Development of Speech-Separation Techniques", Rome Air Development Center, Report RADC-TR-78-105, May 1978.

T.W. Parsons, "Multitalker Separation", Rome Air Develcpment Center, Report RADC-TR-79-242, Oct. 1979.

D.B. Paul, "A Robust Vocoder with Pitch-Adaptive Spectral Envelope Estimation and an Integrated Maximum-Likelihood Pitch Estimator", IEEE ICASSP-79 Conf. Record, pp. 64-68, 1979.

D.B. Paul, "The Spectral Envelope Estimation Vocoder", IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-29, No. 4, pp. 786-794, Aug. 1981.

Y.M. Perlmutter, L.D. Braids, R.H. Frazier, & A.V. Oppenheim, "Evaluation of a Speech Enhancement System", IEEE ICASSP-77 Conf. Record, pp. 212-215, 1977.

T.L. Petersen, "Acoustic Signal Processing in the Context of a Perceptual Model", Ph.D. dissertation, University of Utah, Salt Lake City, 1980.

L.R. Rabiner & R.W. Schafer, Digital Processing of Speech

Signals, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1978.

B. Scharf, "Critical Bands" in Foundations of Modern Auditory Theory, (J.V. Tobias, ed.), Academic Press, Inc., New York, Vol. I, pp. 159-202, 1970.

M.R. Schroeder, "Models of Hearing", Proceedings of the IEEE, Vol. 63, No. 9, pp. 1332-1350, Sept. 1975.

A. Sekey & B.A. Hanson, "Improved One-Bark Bandwidth Auditory Filter", submitted for publication to the J. Acoust. Soc. Amer., 1983.

V.C. Shields, Jr., "Separation of Added Speech Signals by Digital Comb Filtering", S.M. Thesis, Dept. of Electrical Engineering, M.I.T., Cambridge, 1970.

S. Siegel, Nonparametric Statistics for the Behavioral Sciences, McGraw-Hill Book Company, New York, pp. 35-42 & 247, 1956.

A.M. Small, "Psychoacoustics", in Normal Aspects of Speech, Hearing, and Language, ed., F.D. Minifie, T.J. Hixon, & F. Williams, Prentice-Hall, Inc., Englewood Cliffs, N.J., pp. 343-420, 1973.

M.R. Spiegel, Schaum's Outline of Theory and Problems of Statistics, McGraw-Hill Book Company, New York, pp. 143-182 & 343, 1961.

H.W. Strube, "Separation of Several Speakers Recorded By Two Microphones (Cocktail-Party Processing)", Signal Processing, Vol. 3, pp. 355-364, 1981.

J.B. Thomas, An Introduction to Statistical Communication Theory, John Wiley & Sons, Inc., New York, London, Sydney, Toronto, p. 222, 1969.

W.D. Voiers, "Diagnostic Evaluation of Speech Intelligibility", in Speech Intelligibility and Speaker Recognition, (M.E. Hawley, ed.), pp. 374-387, Dowden, Hutchinson & Ross, Inc., 1977.

M.R. Weiss, E. Aschkenasy, & T.W. Parsons, "Study and Development of the INTE Technique for Improving Speech Intelligibility", Nicolet Scientific Corp., Final Rep. NSC-FR/4023, Dec. 1974.

M.R. Weiss & E. Aschkenasy, "Automatic Detection and Enhancement of Speech Signals", Rome Air Development Center, Report RADC-TR-75-77, Mar. 1975.

M.R. Weiss, E. Aschkenasy, & T.W. Parsons, "Study and Development of the INTEL Technique for Improving Speech Intelligibility", Rome Air Development Center, Report RADC-TR-75-108, pp. 26-29, Apr. 1975.

D.Y. Wong, "Evaluation and Improvements to the Quality and Intelligibility of Linear Prediction Voice Coding", Ph.D. Dissertation, Univ. of Calif. Santa Barbara, Mar. 1979.

D.Y. Wong, "Low Bit Rate Encoding/Decoding Algorithm Development", Final Report on Contract MDA904-78-C-0526, Signal Technology, Inc., Santa Barbara, Calif., Aug. 1979.

L.L. Young, Jr. & J. Goodman, "The Effects of Peak Clipping on Speech Intelligibility in the Presence of a Competing Message", IEEE ICASSP-77 Conf. Record, pp. 216-218, 1977.

E. Zwicker, "Subdivision of the Audible Frequency Range into Critical Bands (Frequenzgruppen)", J. Acoust. Soc. Amer., Vol. 33, No. 2, p. 248, Feb. 1961.

END

FILMED

1-84

DTIC